
PLANT PROTEOMICS

Technologies, Strategies, and Applications

Edited by
Ganesh Kumar Agrawal
Randeep Rakwal



WILEY

A JOHN WILEY & SONS, INC., PUBLICATION



PLANT PROTEOMICS



PLANT PROTEOMICS

Technologies, Strategies, and Applications

Edited by
Ganesh Kumar Agrawal
Randeep Rakwal



WILEY

A JOHN WILEY & SONS, INC., PUBLICATION



Copyright © 2008 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Agrawal, Ganesh Kumar.

Plant proteomics : technologies, strategies, and applications / Ganesh Kumar

Agrawal, Randeep Rakwal.

p. cm.

Includes index.

ISBN 978-0-470-06976-9 (cloth)

1. Plant proteins. 2. Plant proteomics. I. Rakwal, Randeep. II. Title.

QK898.P8A37 2008

572'.62—dc22

2007051404

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE	xix
CONTRIBUTORS	xxiii
ACRONYMS AND ABBREVIATIONS	xxix

1 AN INTRODUCTION TO PROTEOMICS: APPLICATIONS TO PLANT BIOLOGY	1
<i>Ralph A. Bradshaw</i>	
1.1 Proteomics Defined	2
1.2 Proteomics Applied	5
References	6

PART I TECHNOLOGIES

2 GEL-BASED PROTEOMICS	11
<i>Pier Giorgio Righetti, Paolo Antonioli, Carolina Simò, and Attilio Citterio</i>	
2.1 Introduction and Brief Bibliographic Review	11
2.2 SDS-PAGE	12
2.3 IEF	17
2.4 2D Maps	22
2.5 Conclusions	28
2.6 Five-Year Viewpoint	29
References	30
3 MASS SPECTROMETRY-BASED PROTEOMICS: IDENTIFYING PLANT PROTEINS	33
<i>Eveline Bergmüller, Sacha Baginsky, and Wilhelm Gruissem</i>	
3.1 Introduction and Brief Bibliographic Review	33
3.2 Instrumentation	34
3.3 MALDI	35

3.4	ESI	36
3.5	Mass Analyzers	36
3.6	Ion Detectors	38
3.7	Sample Preparation	39
3.8	Protein Identification	40
3.9	Conclusions	44
3.10	Five-Year Viewpoint	44
	References	44
4	CHEMICAL PROTEOMICS	47
	<i>Miriam C. Hagenstein, Olaf Kruse, and Norbert Sewald</i>	
4.1	Introduction	47
4.2	Strategies For Activity-Based Protein Profiling (ABPP)	47
4.3	Case Study: Development of Molecular Tools Targeting Plant Kinases	53
4.4	Conclusions	57
4.5	Five-Year Viewpoint	58
	References	58
5	THE <i>ARABIDOPSIS</i> LOCALIZOME: SUBCELLULAR PROTEIN LOCALIZATION AND INTERACTIONS IN <i>ARABIDOPSIS</i>	61
	<i>Georgios Kitsios, Nicolas Tsesmetzis, Max Bush, and John H. Doonan</i>	
5.1	Protein Compartmentalization in Plant Cells	61
5.2	Experimental Determination of Protein Localization	66
5.3	<i>In Vivo</i> Imaging Approaches to Protein Localization and Interaction	69
5.4	Plant Cell Cultures for Studying Protein Localization	73
5.5	Protein–Protein Interaction <i>In Vivo</i> : FRET	74
5.6	Perspectives: Integrating Predictive and Experimental Protein Localization Data	77
	References	77
6	SECRETOME: TOWARD DECIPHERING THE SECRETORY PATHWAYS AND BEYOND	83
	<i>Young-Ho Jung, Ganesh Kumar Agrawal, Randeep Rakwal, and Nam-Soo Jwa</i>	
6.1	Introduction and Brief Bibliographic Review	83
6.2	Methodology and Strategy	86
6.3	A Case Study: <i>In Planta</i> and <i>In Vitro</i> Protein Profiles of Soluble and Secreted Proteins in Rice	87

6.4	Conclusions	87
6.5	Five-Year Viewpoint	89
	References	89

7 PEPTIDOMICS 91

Peter Schulz-Knappe

7.1	Introduction and Brief Bibliographic Review	91
7.2	Separation Technology	93
7.3	MS Technology	95
7.4	Bioinformatics and Data Mining	97
7.5	Differential Peptide Display	98
7.6	ID LC–MALDI	99
7.7	2D CA–RP–LC–ESI–MS	100
7.8	Applications	100
7.9	Peptides and Proteases	101
7.10	Conclusions	102
7.11	Five-Year Viewpoint	103
	References	103

PART II COMPUTATIONAL PROTEOMICS

8 BIOINFORMATICS IN GEL-BASED PROTEOMICS 107

Åsa M. Wheelock and Craig E. Wheelock

8.1	Introduction and Brief Bibliographic Review	107
8.2	Methodology and Strategy	112
8.3	Experimental Results and Applications	120
8.4	Conclusions	122
8.5	Five-Year Viewpoint	123
	References	124

9 BIOINFORMATICS IN MS-BASED PROTEOMICS 127

Jacques Colinge

9.1	Introduction	127
9.2	Database Searching	128
9.3	Peptide <i>De Novo</i> Sequencing	135
9.4	Conclusions and Five-Year Viewpoint	138
	References	139

PART III EXPRESSION PROTEOMICS

10	AN OVERVIEW OF THE <i>ARABIDOPSIS</i> PROTEOME	143
	<i>Jacques Bourguignon and Michel Jaquinod</i>	
10.1	Introduction and Brief Bibliographic Review	143
10.2	Methodology and Strategy	146
10.3	Experimental Results and Applications	147
10.4	Conclusions	160
10.5	Five-Year Viewpoint	160
	References	161
11	RICE PROTEOME AT A GLANCE	165
	<i>Ganesh Kumar Agrawal and Randeep Rakwal</i>	
11.1	Introduction and Brief Bibliographic Review	165
11.2	Methodology and Strategy	167
11.3	Experimental Results and Applications	172
11.4	Conclusions	177
11.5	Five-Year Viewpoint	177
	References	178
12	PROTEOMICS OF LEGUME PLANTS	179
	<i>Satish Nagaraj, Zhentian Lei, Bonnie Watson, and Lloyd W. Sumner</i>	
12.1	Introduction and Brief Bibliographic Review	179
12.2	Proteomics of Model Legume <i>Medicago truncatula</i>	180
12.3	Proteomics of Model System <i>Lotus japonicus</i>	183
12.4	Proteomics of Soybean, <i>Glycine max</i>	183
12.5	Proteomics of Alfalfa, <i>Medicago sativa</i>	184
12.6	Conclusions	186
12.7	Five-Year Viewpoint	187
	References	187
13	PROTEOME OF SEED DEVELOPMENT AND GERMINATION	191
	<i>Julie Catusse, Loïc Rajjou, Claudette Job, and Dominique Job</i>	
13.1	Introduction	191
13.2	Proteomics of Developing Seeds	192
13.3	Proteomics of Mature Seeds	196
13.4	Proteomics of Germinating Seeds	198
13.5	Proteomics of Somatic Embryogenesis	201
13.6	Metabolic Control of Seed Development and Germination	203

13.7	Conclusions	204
13.8	Five-year Viewpoint	205
	References	205
14	ENDOSPERM AND AMYLOPLAST PROTEOMES OF WHEAT GRAIN	207
	<i>William J. Hurkman, William H. Vensel, Frances M. DuPont, Susan B. Altenbach, and Bob B. Buchanan</i>	
14.1	Introduction	207
14.2	Fractionation and Analysis of Endosperm Proteins	208
14.3	The Endosperm Proteome	210
14.4	The Amyloplast Proteome	215
14.5	Conclusions	219
14.6	Five-Year Viewpoint	219
	References	220
15	ROOT PROTEOME	223
	<i>Kuo-Chen Yeh, Chyi-Chuann Chen, and Chuan-Ming Yeh</i>	
15.1	Introduction	223
15.2	Application of Proteomics in the Study of Root Development	225
15.3	Proteome Analysis of Response to Plant Hormones	232
15.4	Proteome Analysis of Gravity Response	233
15.5	Action to Microbe–Plant Symbiosis	233
15.6	Proteome to Cope With Environmental Stresses	234
15.7	Conclusions and Five-Year Viewpoint	236
	References	237
16	LEAF PROTEOME	239
	<i>Bin Kang, Shuyang Tu, Jiyuan Zhang, and Siqi Liu</i>	
16.1	Introduction	239
16.2	Proteomics Studies in Plant Leaf	240
16.3	Conclusions	247
16.4	Five-Year Viewpoint	247
	References	248
17	ANTHER PROTEOME	249
	<i>Nijat Imin</i>	
17.1	Introduction and Brief Bibliographic Review	249
17.2	Methodology and Strategy	251
17.3	Conclusions	258

17.4 Five-Year Viewpoint	258
References	259

18 POLLEN PROTEOME 261

Sandra Noir

18.1 Introduction	261
18.2 Bibliographic Review: Pollen Development and Function	262
18.3 Methodology and Experimental Results	264
18.4 Conclusions	271
18.5 Five-Year Viewpoint	272
References	272

19 MICROTUBULE-BINDING PROTEINS 275

Lori A. Vickerman and Douglas G. Muench

19.1 Introduction	275
19.2 Brief Bibliographic Review	276
19.3 Strategies and Methodologies	277
19.4 Experimental Results and Applications	280
19.5 Conclusions	286
19.6 Five-Year Viewpoint	287
References	287

PART IV ORGANELLE PROTEOMICS

20 CELL WALL 293

Elisabeth Jamet, Hervé Canut, Cécile Albenne, Georges Boudart, and Rafael Pont-Lezica

20.1 Introduction	293
20.2 Brief Bibliographic Review	296
20.3 Specific Methodology and Strategies	296
20.4 Experimental Results and Applications	300
20.5 Conclusions	304
20.6 Five-Year Viewpoint	304
References	306

21 PLASMA MEMBRANE: A PECULIAR STATUS AMONG THE CELL MEMBRANE SYSTEMS 309

Geneviève Ephritikhine, Anne Marmagne, Thierry Meinzel, and Myriam Ferro

21.1 Introduction and Brief Bibliographic Review	309
21.2 Specific Methodologies and Strategies (Box 21.1)	310

21.3	Experimental Results and Applications: Toward an Informative Database of the Plant PM Proteome	319
21.4	Conclusions	323
21.5	Five-Year Viewpoint	324
	References	325
22	NUCLEUS	327
	<i>Subhra Chakraborty, Aarti Pandey, Asis Datta, and Niranjana Chakraborty</i>	
22.1	Introduction and Brief Bibliographic Review	327
22.2	Methodology and Strategy	328
22.3	Experimental Results and Discussion	330
22.4	Conclusions	335
22.5	Five-Year Viewpoint	336
	References	337
23	CHLOROPLAST	339
	<i>Thomas Kieselbach and Wolfgang P. Schröder</i>	
23.1	Introduction	339
23.2	Brief Bibliographic Review	341
23.3	Methodology and Strategy	341
23.4	Experimental Results and Applications	344
23.5	Conclusions	348
23.6	Five-Year Viewpoint	348
	References	349
24	ETIOPLAST	351
	<i>Anne von Zychlinski, Sonja Reiland, Wilhelm Gruissem, and Sacha Baginsky</i>	
24.1	Introduction	351
24.2	Etioplast Proteome Analysis	352
24.3	Conclusions	358
24.4	Five-Year Viewpoint	359
	References	359
25	THE PLANT MITOCHONDRIAL PROTEOME AND THE CHALLENGE OF HYDROPHOBIC PROTEIN ANALYSIS	361
	<i>Yew-Foon Tan and A. Harvey Millar</i>	
25.1	Introduction	361
25.2	Brief Bibliographic Review	363
25.3	Methodology and Strategy	364
25.4	Experimental Results and Applications	369

25.5 Conclusions	373
25.6 Five-Year Viewpoint	373
References	374

26 PEROXISOME 377

Yuko Arai, Youichiro Fukao, Makoto Hayashi, and Mikio Nishimura

26.1 Introduction	377
26.2 Brief Bibliographic Review	379
26.3 Specific Methodologies and Strategies	379
26.4 Experimental Results and Applications	383
26.5 Conclusions	386
26.6 Five-Year Viewpoint	387
References	388

27 UNRAVELING PLANT VACUOLES BY PROTEOMICS 391

Songqin Pan and Natasha Raikhel

27.1 Introduction	391
27.2 Bibliographic Review: Role of Proteomics in Understanding Biology of Plant Vacuoles	393
27.3 Methodology and Strategies	395
27.4 A Case-Study of Label-Free Quantitative Proteomics	401
27.5 Conclusions	402
27.6 Five-Year Viewpoint	403
References	404

28 OIL BODIES 407

Pascale Jolivet, Luc Negroni, Sabine d'Andréa, and Thierry Chardot

28.1 Introduction and Brief Bibliographic Review	407
28.2 Methodology and Strategy	408
28.3 Conclusions	413
28.4 Five-Year Viewpoint	414
References	416

PART V MODIFICATION PROTEOMICS

29 PHOSPHOPROTEINS: WHERE ARE WE TODAY? 421

Florian Wolschin and Wolfram Weckwerth

29.1 Introduction	421
29.2 Phosphoprotein and Phosphopeptide Enrichment	422

29.3	Detection of Protein Phosphorylation and Determination of Phosphorylation Sites	427
29.4	MS-Based Approaches	432
29.5	Biological Implications of Protein Multisite Phosphorylation in Plants	434
29.6	Large-Scale Proteomics Studies	437
29.7	Conclusions and Five-Year Viewpoint	438
	References	438
30	PROTEOME ANALYSIS OF THE UBIQUITIN PATHWAY	443
	<i>Junmin Peng</i>	
30.1	Introduction	443
30.2	Brief Bibliographic Review	444
30.3	Specific Methodologies and Strategies	445
30.4	Experimental Results and Applications	448
30.5	Conclusions	451
30.6	Five-Year Viewpoint	452
	References	452
31	ANALYSIS OF THE N-GLYCOSYLATION OF PROTEINS IN PLANTS	455
	<i>Willy Morelle</i>	
31.1	Introduction	455
31.2	Brief Bibliographic Review	456
31.3	Methodology and Strategy	458
31.4	Conclusions	465
31.5	Five-Year Viewpoint	466
	References	467
32	FUNCTIONAL ANALYSIS AND PHOSPHORYLATION SITE MAPPING OF LEUCINE-RICH REPEAT RECEPTOR-LIKE KINASES	469
	<i>Steven D. Clouse, Michael B. Goshe, Steven C. Huber, and Jia Li</i>	
32.1	Introduction and Brief Bibliographic Review	469
32.2	Specific Methodologies and Strategies	473
32.3	Experimental Results and Applications	477
32.4	Conclusions and Five-Year Viewpoint	481
	References	482
33	TIME TO SEARCH FOR PROTEIN KINASE SUBSTRATES	485
	<i>Birgit Kersten</i>	
33.1	Introduction and Brief Bibliographic Review	485
33.2	Methodology and Strategy	486

33.3 Experimental Results on Plant MAPK Downstream Signaling and Applications	494
33.4 Conclusions	496
33.5 Five-Year Viewpoint	496
References	497

34 TYROSINE PHOSPHORYLATION IN PLANTS: EMERGING EVIDENCE 499

Andrea Carpi, Valeria Rossi, and Francesco Filippini

34.1 Introduction	499
34.2 Brief Bibliographic Review	502
34.3 Methodology and Strategy	505
34.4 Experimental Results and Applications	508
34.5 Conclusions	510
34.6 Five-Year Viewpoint	511
References	512

35 14-3-3 PROTEINS: REGULATORS OF KEY CELLULAR FUNCTIONS 515

Peter C. Morris

35.1 Introduction and Brief Bibliographic Review	515
35.2 Methodology and Strategy	518
35.3 Experimental Results and Applications	519
35.4 Conclusions	522
35.5 Five-Year Viewpoint	522
References	523

PART VI MULTIPROTEIN COMPLEX

36 TAP-TAGGING SYSTEM IN RICE FOR PROTEIN COMPLEX ISOLATION 527

Jai S. Rohila and Michael E. Fromm

36.1 Introduction	527
36.2 Methodolgy and Strategy	528
36.3 Experimental Results and Applications	535
36.4 Conclusions	540
36.5 Five-Year Viewpoint	541
References	541

37	TAP STRATEGY IN <i>ARABIDOPSIS</i> PROTEIN COMPLEX ISOLATION	543
	<i>Vicente Rubio and Xing Wang Deng</i>	
37.1	Introduction and Brief Bibliographic Review	543
37.2	Specific Methodology and Strategies	546
37.3	Experimental Results and Applications	552
37.4	Conclusions	554
37.5	Five-Year Viewpoint	555
	References	555
38	BLUE-NATIVE PAGE IN STUDYING PROTEIN COMPLEXES	557
	<i>Holger Eubel and A. Harvey Millar</i>	
38.1	Introduction	557
38.2	Brief Bibliographic Review	560
38.3	Methodology and Strategy	561
38.4	Experimental Results and Applications	566
38.5	Conclusions	567
38.6	Five-Year Viewpoint	567
	References	568
39	PROTEIN-PROTEIN INTERACTION MAPPING IN PLANTS	571
	<i>Joachim F. Uhrig</i>	
39.1	Introduction and Brief Bibliographic Review	571
39.2	Methodology and Strategy	574
39.3	Experimental Results and Applications	578
39.4	Conclusions	580
39.5	Five-Year Viewpoint	581
	References	581

PART VII PLANT DEFENSE AND STRESS

40	PROTEOMICS IN PLANT DEFENSE RESPONSE	587
	<i>Sun Tae Kim and Kyu Young Kang</i>	
40.1	Introduction	587
40.2	Brief Bibliographic Review	588
40.3	Methodology and Strategies	589
40.4	Experimental Results and Applications	595
40.5	Conclusions	601

40.6 Five-Year Viewpoint	602
References	603
41 PROTEOME ANALYSIS OF CELLULAR RESPONSES TO ABIOTIC STRESSES IN PLANTS	605
<i>Hans-Peter Mock and Andrea Matros</i>	
41.1 Introduction and Brief Bibliographic Review	605
41.2 Summary of Previous Research on Abiotic Stresses in Plants	606
41.3 Comparison of Transcriptomics and Proteomics Data in Analysis of Plant Defense Responses Against Abiotic Stress Factors	621
41.4 Conclusions	624
41.5 Five-Year Viewpoint	624
References	625
42 PROTEOMICS OF BIOTROPHIC PLANT–MICROBE INTERACTIONS: SYMBIOSES LEAD THE MARCH	629
<i>Ghislaine Recorbet and Eliane Dumas-Gaudot</i>	
42.1 Introduction	629
42.2 AM and RNF Symbioses: No Equal Footing for Proteomics	631
42.3 A 2005–2006 Update of the Contribution of Proteomics to AM Symbiosis	634
42.4 A 2005–2006 Update of the Contribution of Proteomics to RNF Symbiosis	637
42.5 Conclusions	640
42.6 Five-Year Viewpoint	640
References	642
43 PROTEOMICS APPROACHES TO CONSTRUCT CALCIUM SIGNALING NETWORKS IN PLANTS	645
<i>Irene S. Day and A.S.N. Reddy</i>	
43.1 Introduction and Brief Bibliographic Review	645
43.2 Methodology and Experimental Results	649
43.3 Conclusions	656
43.4 Five-Year Viewpoint	656
References	657
PART VIII STRUCTURAL PROTEOMICS	
44 CELL-FREE EXPRESSION SYSTEM FOR EUKARYOTIC PROTEINS	661
<i>Yaeta Endo and Tatsuya Sawasaki</i>	
44.1 Introduction	661

44.2	Development of Wheat Germ Cell-Free Protein Synthesis System	662
44.3	Application to Functional Proteomics	663
44.4	Application to Structural Proteomics	665
44.5	Conclusions and Five-Year Viewpoint	667
	References	668
45	PROTEIN STRUCTURE DETERMINATION	671
	<i>Jian-Hua Zhao and Hsuan-Liang Liu</i>	
45.1	Introduction	671
45.2	Instrumental Methods for Protein Structure Determination	672
45.3	Computational Approaches for Protein Structure Prediction	678
45.4	Conclusions	687
45.5	Five-Year Viewpoint	687
	References	689
 PART IX OTHER TOPICS IN PLANT PROTEOMICS		
46	PROTEOMICS IN CONTEXT OF SYSTEMS BIOLOGY	695
	<i>Serhiy Souchelnyskyi</i>	
46.1	Introduction	695
46.2	What Systems Biology Requires	696
46.3	What Proteomics Provides	699
46.4	Representation of Information About Proteins and Their Integration into Systems Biology Tools	706
46.5	Conclusions	708
46.6	Five-Year Viewpoint	708
	References	709
47	PROTEOMICS IN DEVELOPING COUNTRIES	713
	<i>Nat N. V. Kav, Sanjeeva Srivastava, William Yajima, and Shakir Ali</i>	
47.1	Introduction	713
47.2	Proteome Research in the Developing World	719
47.3	Conclusions	725
47.4	Five-Year Viewpoint	726
	References	726
	INDEX	731

PREFACE

The beginning of the twenty-first century is certainly a great time to be involved in plant proteomics. This new millennium has placed an ever growing amount of sophisticated technology (i.e., the ever growing list of annotated genes and genome sequence databases from *Arabidopsis*, rice and other plant species) at the disposal of the modern scientist to the benefit of all. The research performed with this technology has the potential to bring the answers to important and complex biological questions and problems (especially those relating to crop plants and the human food supply) within reach.

Proteomics in plants began in 1990s, but has accelerated with an unexpected pace and momentum since 1999 when many economical techniques (which are standard today) were developed to characterize proteins on a proteome-wide scale. The proteomics burst seen in the literature in recent years has made a significant impact on plant biology mainly by answering many of the questions associated with the genome annotation and the number of functional proteins expressed in a given organism. Indeed, when one looks at the progress achieved to-date in the field of plant proteomics and its overall impact on biological research, it is clear just how essential proteomics is to our understanding of the physiology of any organism. For example, proteomics is perhaps the only means that enables one to fully understand post-translational modifications of proteins. Thus, it would not be far from the truth to say that the “power of proteomics” (and indeed the omic sciences as a whole) is one of the driving forces of twenty-first century biological science.

A current (early 2008) literature survey in PubMed indicates that the number of publications containing the term “plant proteomics” is 100-fold higher today, compared to same period in 1999. Not surprisingly, this survey also indicates that proteomic studies have primarily been conducted in the two widely accepted flowering model plants, the weed *Arabidopsis thaliana* and the cereal crop rice (*Oryza sativa* L.). The publication of the genome sequences of these two plants represents a significant landmark in the history of plant biology. However, one must ask: Are we fully aware of the true potential of plant proteomics and if so, are we using this knowledge to its full effect? The principles of good science are as true in this age of omics still hold true today and the disciplined scientist must keep these principles in mind to avoid rushing blindly into the field (intentionally or unintentionally) without first obtaining a thorough understanding of its fundamental principles.

When one looks at the impressive progress of proteomics in plant science, as well as its immense importance in biological sciences as a whole, it is clear that a need for

a textbook, exists to translate/disseminate the knowledge acquired by leading experts in the field to the wider scientific community. This was the impetus for the book you are currently reading. Though we knew that such a project would be a daunting challenge, we also knew that it would bring the opportunity to work closely with the leading experts of the field. What we did not fully appreciate when we started was how much of a truly unparalleled experience it would be to work with each and every one of the contributors of this book, whom we genuinely thank for being part of this ambitious endeavor.

This book is composed of 9 sections in the following order: overview of proteomics in plant biology, technologies, computational/expression/organelle/modification proteomics, multiprotein complex, plant defense and stress, structural proteomics, systems biology, and proteomics in developing countries. The 47 chapters provide excellent coverage of almost all the studies conducted to-date on plant growth and development at the proteomics level. Each chapter also contains a five-year viewpoint which discusses the scope for investigating proteomes and innovative improvements in proteomic technologies over the next decade. We hope this book will be beneficial in scope and practical knowledge to readers, whose response will be the final judge of the validity of the work.

As a final point, it is fitting that we acknowledge the people who gave their unconditional and inspiring support, without which this book would not have reached completion. First and foremost we would like to thank Professor Dominic M. Desiderio (Department of Neurology, University of Tennessee, Memphis, Tennessee, USA) for being our mentor in this endeavor and enabling completion of this tremendous milestone in our lives.

Secondly, we wish to thank our colleagues and collaborators around the world with whom we have struggled to do “good science,” forming new partnerships and friendships in the process. There is not room to mention all those who have had an effect on us here but Masami Yonekura (Ibaraki University, Japan), Shigeru Tamogami (Akita Prefectural University, Japan), Akihiro Kubo (National Institute of Environmental Sciences, Japan), Nam-Soo Jwa (Sejong University, Korea), Oksoo Han (Chonnam National University, Korea), Birgit Kersten (Max Planck Institute for Molecular Plant Physiology, Germany), Yu Sam Kim and Hyung Wook Nam (Yonsei University, Korea), Hirohiko Hirochika (National Institute of Agrobiological Sciences, Japan), Shoshi Kikuchi (National Institute of Agrobiological Sciences, Japan), Oliver A.H. Jones (University of Cambridge, United Kingdom), and Yoshinori Masuo and Hitoshi Iwahashi (National Institute of Advanced Industrial Science and Technology, Japan) all deserve both mention and appreciation. We would especially like to thank Professor Vishwanath Prasad Agrawal (RLABB, Kathmandu, Nepal) for his directions and guidance in our research. (This is especially true for Ganesh who started his research under Professor Vishwanath’s watchful eyes).

Thirdly, we thank the Editorial Team (Scientific, Technical, Medical, and Scholarly Division) at John Wiley & Sons, Inc., especially Executive Editor Bob Esposito, Senior Editorial Assistant Brendan Sullivan, Senior Designer Daniel Timek, and, last but not the least, Senior Production Editor Lisa Morano Van Horn for their professional support and patience with our queries and correspondence.

Finally, to this long list of supporters, we must add our thanks for the personal sacrifices by our families, especially our wives and children. Randeep's wife Junko Shibato also contributed greatly to the technical aspects of the book working alongside with him in the laboratory. Our parents who brought us into this world and who taught and inspired us to contribute to society and do our duty also deserve special mention. To you the reader we also extend our thanks and appreciation. We hope this work will be useful to you.

GANESH KUMAR AGRAWAL
RANDEEP RAKWAL

Kathmandu, Nepal
Tsukuba, Japan
May 2008

CONTRIBUTORS

Ganesh Kumar Agrawal, Research Laboratory for Biotechnology and Biochemistry (RLABB), Kathmandu, Nepal

Cécile Albenne, Surface Cellular and Signalization for Plants, Projects in Plant Biotechnology, UMR, CNRS, Paul Sabatier-Toulouse III University, Castanet-Tolosan, France

Shakir Ali, Department of Biochemistry, Hamdard University, Hamdard Nagar, New Delhi, India

Susan B. Altenbach, U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Albany, California

Paolo Antonioli, Department of Chemistry, Materials and Chemical Engineering “Giulio Natta,” Polytechnic of Milan, Milan, Italy

Yuko Arai, National Institute for Basic Biology, Myodaiji, Okazaki, Japan

Sacha Baginsky, Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland

Eveline Bergmüller, Institute of Plant Sciences, ETH Zurich, Switzerland

Georges Boudart, Surface Cellular and Signalization for Plants, Projects in Plant Biotechnology, UMR, CNRS, Paul Sabatier-Toulouse III University, Castanet-Tolosan, France

Jacques Bourguignon, Laboratory of Cellular Plant Physiology, Institute of Research and Technologies for Life Sciences, CNRS, UJF, INRA, CEA, Grenoble, France

Ralph A. Bradshaw, Mass Spectrometry Facility, University of California, San Francisco, California

Bob B. Buchanan, Department of Plant and Microbial Biology, University of California, Berkeley, California

Max Bush, John Innes Centre, Norwich, England

Hervé Canut, Surface Cellular and Signalization for Plants, Projects in Plant Biotechnology, UMR, CNRS, Paul Sabatier-Toulouse III University, Castanet-Tolosan, France

Andrea Carpi, Department of Biological Chemistry, University of Padua, Padua, Italy

Julie Catusse, CNRS/UCBL/INSA/Bayer CropScience Joint Laboratory, Bayer CropScience, Lyon, France

Niranjan Chakraborty, National Centre for Plant Genome Research, JNU Campus, New Delhi, India

Subhra Chakraborty, National Centre for Plant Genome Research, JNU Campus, New Delhi, India

Thierry Chardot, UMR 206 Chemical Biology, INRA, INA-PG, Thiverval Grignon, France

Chyi-Chuann Chen, Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan

Attilio Citterio, Department of Chemistry, Materials and Chemical Engineering “Giulio Natta,” Polytechnic of Milan, Milan, Italy

Steven D. Clouse, Department of Horticultural Science, North Carolina State University, Raleigh, North Carolina

Jacques Colinge, Ce-M-M Research Center for Molecular Medicine, Austrian Academy of Science, Director’s Group, Vienna, Austria

Sabine d’Andréa, UMR 206 Chemical Biology, INRA, INA-PG, Thiverval Grignon, France

Asis Datta, National Center for Plant Genome Research, JNU Campus, New Delhi, India

Irene S. Day, Department of Biology, Programs in Molecular Plant Biology, and Cell and Molecular Biology, Colorado State University, Fort Collins, Colorado

Xing Wang Deng, Department of Molecular, cell, and Development Biology, Yale University, New Haven, Connecticut

John H. Doonan, John Innes Centre, Norwich, England

Eliane Dumas-Gaudot, UMR 1088 INRA/CNRS 5184/UB Plant Microbe Environment, INRA/CMSE, Dijon, France

Frances M. DuPont, U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Albany, California

Yaeta Endo, Cell-Free Science and Technology Research Center, Ehime University, Matsuyama, Japan

Geneviève Ephritikhine, National Center of Scientific Research, Institute of Plant Sciences, CNRS, UPR, Gif sur Yvette, France and UFR Natural Biology Sciences, University of Paris, Paris, France

Holger Eubel, Centre of Excellence in Plant Energy Biology, School of Biomedical and Chemical Sciences/ARC, University of Western Australia, Crawley, Australia

Myriam Ferro, Laboratory for the Study of Dynamic Proteomics, INSERM, IRTSV, CEA-DSV, IUJF, Grenoble, France

Francesco Filippini, Department of Biology, University of Padua, Padua, Italy

Michael E. Fromm, Plant Science Initiative, University of Nebraska, Lincoln, Nebraska

Youichiro Fukao, Nara Institute of Science and Technology, Takayama, Ikoma, Japan

- Michael B. Goshe**, Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, North Carolina
- Wilhelm Gruissem**, Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland
- Miriam C. Hagenstein**, Department of Chemistry, Bielefeld University, Bielefeld, Germany
- Makoto Hayashi**, National Institute for Basic Biology, Myodaiji, Okazaki, Japan
- Steven C. Huber**, USDA/ARS, University of Illinois, Urbana, Illinois
- William J. Hurkman**, U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Albany, California
- Nijat Imin**, ARC Center of Excellence for Integrative Legume Research, Genomic Interactions Group, Research School of Biological Sciences, The Australian National University, Canberra, Australia
- Elisabeth Jamet**, Surface Cellular and Signalization for Plants, Projects in Plant Biotechnology, UMR, CNRS, Paul Sabatier-Toulouse III University, Castanet-Tolosan, France
- Michel Jaquinod**, Laboratory of Cellular Plant Physiology, Institute of Research and Technologies for Life Sciences, CNRS, UJF, INRA, CEA, Grenoble, France
- Claudette Job**, CNRS/UCBL/INSA/Bayer CropScience Joint Laboratory, Bayer CropScience, Lyon, France
- Dominique Job**, CNRS/UCBL/INSA/Bayer CropScience Joint Laboratory, Bayer CropScience, Lyon, France
- Pascale Jolivet**, UMR 206 Chemical Biology, INRA, INA-PG, Thiverval Grignon, France
- Young-Ho Jung**, Department of Molecular Biology, Sejong University, Gunja-dong, Seoul, South Korea
- Nam-Soo Jwa**, Department of Molecular Biology, Sejong University, Gunja-dong, Seoul, South Korea
- Bin Kang**, Center of Proteomic Analysis, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China
- Kyu Young Kang**, Environmental Biotechnology National Core Research Center, Division of Applied Life Science, Gyeongsang National University, Jinju, Korea
- Nat N. V. Kav**, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Canada
- Birgit Kersten**, Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany
- Thomas Kieselbach**, Department of Chemistry, Umeå University, Umeå, Sweden
- Sun Tae Kim**, Environmental Biotechnology National Core Research Center, Division of Applied Life Science, Gyeongsang National University, Jinju, Korea
- Georgios Kitsios**, John Innes Centre, Norwich, England
- Olaf Kruse**, Department of Biology, Bielefeld University, Bielefeld, Germany

Zhentian Lei, Biological Mass Spectrometry, Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma

Jia Li, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma

Hsuan-Liang Liu, Department of Chemical Engineering and Biotechnology, National Taipei University of Technology, Taipei, Taiwan

Siqi Liu, Center of Proteomic Analysis, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China

Anne Marmagne, National Center of Scientific Research, Institute of Plant Sciences, CNRS, UPR, Gif sur Yvette, France

Andrea Matros, Leibniz-Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

Thierry Meinzel, National Center of Scientific Research, Institute of Plant Sciences, CNRS, UPR, Gif sur Yvette, France

A. Harvey Millar, ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Australia

Hans-Peter Mock, Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

Willy Morelle, Structural Glycobiology Unit, Structure and Function Unit for Mixed Recherche, CNRS/USTL 8576, University of Science and Technology of Lille, Villeneuve d'Ascq, France

Peter C. Morris, School of Life Sciences, Heriot-Watt University, Riccarton, Edinburgh, United Kingdom

Douglas G. Muench, Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada

Satish Nagaraj, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma

Luc Negroni, Faculty of Medicine, Proteomic Analysis Facility, Nice, France

Mikio Nishimura, National Institute for Basic Biology, Myodaiji, Okazaki, Japan

Sandra Noir, Department of Plant–Microbe Interactions, Max Planck Institute for Plant Breeding Research, Köln, Germany

Songqin Pan, Institute for Integrative Genome Biology, Center for Plant Cell Biology and Department of Botany and Plant Sciences, University of California, Riverside, California

Aarti Pandey, National Centre for Plant Genome Research, JNU Campus, New Delhi, India

Junmin Peng, Department of Human Genetics, Center for Neurodegenerative Disease, School of Medicine, Emory University, Atlanta, Georgia

Rafael Pont-Lezica, Surface Cellular and Signalization for Plants, Projects in Plant Biotechnology, UMR, CNRS, Paul Sabatier-Toulouse III University, Castanet-Tolosan, France

- Natasha Raikhel**, Institute for Integrative Genome Biology, Center for Plant Cell Biology and Department of Botany and Plant Sciences, University of California, Riverside, California
- Loïc Rajjou**, UMR 204 INRA/AgroParisTech, Seed Biology Laboratory, Paris France
- Randeep Rakwal**, Research Laboratory for Agricultural Biotechnology and Biochemistry (RLABB), Kathmandu, Nepal and Health Technology, Research Center (HTRC), National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba West, Ibaraki, Japan
- Ghislaine Recorbet**, UMR 1088 INRA/CNRS 5184/UB Plant Microbe Environnement, INRA/CMSE, Dijon, France
- A. S. N. Reddy**, Department of Biology, Programs in Molecular Plant Biology and Cell and Molecular Biology, Colorado State University, Fort Collins, Colorado
- Sonja Reiland**, Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland
- Pier Giorgio Righetti**, Department of Chemistry, Materials and Chemical Engineering “Giulio Natta,” Polytechnic of Milan, Milan, Italy
- Jai S. Rohila**, Department of Plant Pathology, Pennsylvania State University, University Park, Pennsylvania
- Valeria Rossi**, Department of Biology, University of Padua, Padua, Italy
- Vicente Rubio**, National Center for Biotechnology—CSIC, Cantoblanco, Madrid, Spain
- Tatsuya Sawasaki**, Cell-Free Science and Technology Research Center, Ehime University, Matsuyama, Japan
- Wolfgang P. Schröder**, Department of Chemistry, Umeå University, Umeå, Sweden
- Peter Schulz-Knappe**, Proteome Sciences R&D GmbH & Co., Frankfurt Germany
- Norbert Sewald**, Department of Chemistry, Bielefeld University, Bielefeld, Germany
- Carolina Simò**, Department of Chemistry, Materials and Chemical Engineering Giulio Natta, Polytechnic of Milan, Milan, Italy
- Serhiy Souchelnytskyi**, Medical Proteomics Group, Karolinska Biomics Center, Karolinska University Hospital, Stockholm, Sweden
- Sanjeeva Srivastava**, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Canada
- Lloyd W. Sumner**, Biological Mass Spectrometry, Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma
- Yew-Foon Tan**, ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Australia
- Nicolas Tsismetzis**, John Innes Centre, Norwich, England
- Shuyang Tu**, Center of Proteomic Analysis, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China
- Joachim F. Uhrig**, Botanical Institute III, University of Köln, Köln, Germany

William H. Vensel, U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Albany, California

Lori A. Vickerman, Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada

Anne von Zychlinski, Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland

Bonnie Watson, Biological Mass Spectrometry, Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma

Wolfram Weckwerth, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Åsa M. Wheelock, Karolinska Biomics Center and Department of Medicine, Division of Respiratory Medicine, Karolinska Institute, Stockholm, Sweden

Craig E. Wheelock, Karolinska Biomics Center, and Department of Medical Biochemistry and Biophysics, Division of Physiological Chemistry II, Karolinska Institute, Stockholm, Sweden

Florian Wolschin, Arizona State University, School of Life Sciences, Tempe, Arizona

William Yajima, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Canada

Chuan-Ming Yeh, Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan

Kuo-Chen Yeh, Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan

Jiyuan Zhang, Center of Proteomic Analysis, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China

Jian-Hua Zhao, Department of Chemical Engineering and Biotechnology, National Taipei University of Technology, Taipei, Taiwan

ACRONYMS AND ABBREVIATIONS

α	alpha
ABA	abscisic acid
ABC-HRP	avidin:biotinylated enzyme complex–horseradish peroxidase
ABP	affinity-based probes
ABP1	auxin-binding protein 1
ABPP	activity-based protein(ome) profiling
<i>A. brassicicola</i>	<i>Alternaria brassicicola</i>
ABRC	<i>Arabidopsis</i> biological resource center
ACC	1-aminocyclopropane-1-carboxylate
ACS	1-aminocyclopropane-1-carboxylic acid synthase
ACX	acyl-CoA oxidase
AD	activation domain
ADF	actin depolymerization factor
ADH	alcohol dehydrogenase
<i>ad hoc</i>	for the specific purpose, case, or situation at hand and for no other
ADP	adenosine diphosphate
AdoHcy	S-adenosylhomocysteine
AdoMet	S-adenosylmethionine
<i>A. euteiches</i>	<i>Aphanomyces euteiches</i>
AGP	arabinogalactan proteins
AGPase	ADP-glucose pyrophosphorylase
AGML	annotated gel markup language
<i>A. hypogaea</i>	<i>Arachis hypogaea</i>
Ala	alanine
ALDH	aldehyde dehydrogenase
<i>A. lesbiacum</i>	<i>Alyssum lesbiacum</i>
ALIS	alexa-fluor-labeled internal protein standard
Al(OH) ₃	aluminum hydroxide
AM	arbuscular mycorrhiza(al)
AMAs	antibody microarrays
AMPL	<i>Arabidopsis</i> membrane protein library

AMPS	2-acrylamido-2-methyl propane sulphonic acid
AMRT	accurate mass and retention time
ANN	artificial neural network
ANTS	8-aminonaphthalene-1,3,6-trisulfonic acid
AOA	aminooxyacetic acid
AOX	alternate oxidase
AP	alkaline phosphatase
APC	anaphase-promoting complex
APS	ammonium persulfate
APX	ascorbate peroxidase
AQUA	absolute quantification of proteins
Ar	argon
Arg	arginine
As	arsenic
Asn	asparagine
Asp	aspartate/aspartic acid
ASR	ABA-stress-ripening
ATCase	aspartate carbamoyltransferase
<i>A. thaliana</i>	<i>Arabidopsis thaliana</i>
ATP	adenosine triphosphate
<i>A. tumefaciens</i>	<i>Agrobacterium tumefaciens</i>
<i>A. victoria</i>	<i>Aequorea victoria</i>
β	beta
B	boron
Ba	barium
BAC	benzyltrimethyl- <i>n</i> -hexadecylammonium chloride
BAK1	BRI1-associated receptor kinase
<i>B. cinerea</i>	<i>Botrytis cinerea</i>
BCIP	5-bromo-4-chloro-3'-indolylphosphate p-toluidine salt
Bcp	blue copper-binding protein
BEMAD	β -elimination followed by Michael addition with dithiothreitol
BiFC	bimolecular fluorescence complementation
BiP	binding protein
<i>B. japonicum</i>	<i>Bradyrhizobium japonicum</i>
BL	brassinolide
<i>blm</i>	blast lesion mimic
<i>B. napus</i>	<i>Brassica napus</i>
BN-PAGE	blue native-polyacrylamide gel electrophoresis
BP	band pass
Bpa	benzoylphenylalanine
BR	brassinosteroid
BRET	bioluminescence resonance energy transfer
BRI1	brassinosteroid insensitive 1
BSA	bovine serum albumin
<i>B. subtilis</i>	<i>Bacillus subtilis</i>

<i>C. annuum</i>	<i>Capsicum annuum</i>
<i>C. arietinum</i>	<i>Cicer arietinum</i>
<i>ca.</i>	<i>circa</i> (about)
CA	carrier ampholyte
Ca ²⁺	calcium ion
CaCl ₂	calcium chloride
CADD	computer-aided drug design
CA-IEF/IPG	carrier ampholyte isoelectric focusing/immobilized pH gradients
CaM	calmodulin
cAMP	cyclic adenosine monophosphate
CaMV	cauliflower mosaic virus
CAX	cation exchange chromatography
CBB	Coomassie Brilliant Blue
CBDB	CaM-binding buffer
CBL	calcein B-like
CBP	CaM-binding protein/peptide
CCD	charge couple device
CCOMT	caffeoyl-CoA 3- <i>O</i> -methyltransferase
Cd	cadmium
CD	circular dichroism
CDA	covariant discriminant algorithm
CDC	cell division cycle
CDK	cyclin-dependent (protein) kinase
cDNA	complementary deoxyribonucleic acid
CDPK	calcium-dependent protein kinase
CDTA	cyclohexanediaminetetraacetic acid
Ce	cesium
CEB	CaM-elution buffer
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
<i>C. ensiformi</i>	<i>Canavalia ensiformi</i>
CFCF	continuous-flow cell-free
CFP	cyan fluorescent protein
CHAPS	3-[(3-choramidopropyl)dimethylamino]propane-1-sulfonate
CHR	chalcone reductase
CHS	chalcone synthase
CI	chemical ionization
CID	collision-induced dissociation
CK	casein kinase
C/M	chloroform/methanol
CML	CaM-like
CMS	cytoplasmic male sterility
<i>C. nucifera</i>	<i>Cocos nucifera</i>
CN-PAGE	colorless- or clear-native PAGE
CO ₂	carbon dioxide
COFRADIC	combined fractional diagonal chromatography

COMT	caffeate 3- <i>O</i> -methyltransferase
ConA	concanavalin A
COP	constitutive photomorphogeneic
COSY	correlated spectroscopy
COX	cytochrome <i>c</i> oxidase
CPY	carboxypeptidase Y
Cr	chromium
<i>C. reinhardtii</i>	<i>Chlamydomonas reinhardtii</i>
CRK	CDPK-related kinase
CRT	C-repeat
CS	citrate synthase
CSN	COP9 signalosome
CTAB	cetyltrimethylammonium bromide
CTPP	carboxy-terminal propeptide
Cu	copper
CV	coefficient of variance
CWP	cell wall proteins
Cy	cyanine
CYCB	cyclin B
CYP	cyclophilin
Cys	cysteine
CZE	capillary zone electrophoresis
D	dormant
Da	dalton
DAB	3,3'-diaminobenzidine tetrahydrochloride
DAGAT	diacylglycerol acyltransferase
DAP	days after pollination
DBD	DNA-binding domain
<i>D. carota</i>	<i>Daucus carota</i>
DDA	data-dependent acquisition
DDM	<i>n</i> -dodecylmaltoside
DE	delayed ion extraction
DHAR	dehydroascorbate reductase
DHB	2,5-dihydroxybenzoic acid
Dicot	dicotyledonous
DIGE	difference in-gel electrophoresis
<i>D. melanogaster</i>	<i>Drosophila melanogaster</i>
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
<i>D. purpurea</i>	<i>Digitalis purpurea</i>
DPA	days post-anthesis
DPTPS	deep purple total protein stain
DRM	detergent-resistant membrane
DSK	dual-specificity kinase
DSP	dual-specificity phosphatase

DTT	dithiothreitol
DUBs	deubiquitinating enzymes
EB	end binding
ECD	electron capture dissociation
<i>E. cichoracearum</i>	<i>Erysiphe cichoracearum</i>
ECL	enhanced chemiluminescence
ECM	extracellular matrix
<i>E. coli</i>	<i>Escherichia coli</i>
EDT	1,2-ethanedithiol
EDTA	ethylenediaminetetraacetic acid
<i>E. elongatum</i>	<i>Elymus elongatum</i>
EES	excitation-emission spectroscopy
EF	elongation factor
EGF	epidermal growth factor
EGTA	ethylene glycol-bis-(beta-aminoethyl ether)- <i>N,N,N',N'</i> -tetraacetic acid
<i>E. histolytica</i>	<i>Entamoeba histolytica</i>
EI	electron ionization
eIF	eukaryotic initiation factor
EM	electron microscope, electron microscopy
EMS	ethyl methanesulfonate
<i>E. pisi</i>	<i>Erysiphe pisi</i>
ER	endoplasmic reticulum
ERAD	endoplasmic-reticulum-associated degradation
ERK	extracellular-signal regulated kinase
ESI	electrospray ionization
ESI–MS/MS	electrospray ionization tandem mass spectrometry
EST	expressed sequence tag
ETD	electron transfer dissociation
FAB	fast atom bombardment
FACE	fluorophore-assisted carbohydrate electrophoresis
FBA	fructose biphosphatase aldolase
FBPase	fructose 1,6-biphosphatase
Fe	iron
Fe ₃ O ₄	iron oxide
<i>F. graminearum</i>	<i>Fusarium graminearum</i>
FeSO ₄	iron sulfate
FHA	forkhead associated
FLIM	fluorescence lifetime imaging microscopy
<i>F. moniliforme</i>	<i>Fusarium moniliforme</i>
FMN	flavin mononucleotide
FN	false negative
FP	false positive
FPF	false-positive fraction
FPLC	fast protein liquid chromatography

FRET	fluorescence resonance energy transfer
FROC	free-response operator characteristics
FSPIM	fluorescence spectral microscopy
FTFLP	fluorescent tagging of full-length proteins
FTICR	Fourier transform ion cyclotron resonance
FTIR	Fourier transform infrared
Ga ³⁺	gallium
GAs	gibberellins
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GAPs	GPI-anchored proteins
GC	gas chromatography
GFP	green fluorescent protein
GGT	glutamate:glyoxylate aminotransferase
<i>G. intraradices</i>	<i>Glomus intraradices</i>
Gln	glutamine
Glu	glutamate/glutamic acid
GluTR	glutamyl-tRNA reductase
Gly	glycine
<i>G. max</i>	<i>Glycine max</i>
<i>G. mosseae</i>	<i>Glomus mosseae</i>
GO	gene ontology
GPI	glycosylphosphatidylinositol
GPX	glutathione peroxidase
GR	glutathione reductase
GRAVY	grand average of hydropathicity
GRPs	glycine-rich RNA-binding proteins
GSAAT	glutamate-1-semialdehyde aminotransferase
GSH	glutathione
GSNO	<i>S</i> -nitrosoglutathione
GST	glutathione <i>S</i> -transferase
GTP	guanosine triphosphate
GUS	β -glucuronidase
GVG	glucocorticoid receptor
<i>G. zea</i>	<i>Gibberella zea</i>
H_1	alternative hypothesis
H_0	null hypothesis
HB	homogenization buffer
HCl	hydrochloric acid
He	helium
HF	hydrogen fluoride
His	histidine
HMM	hidden Markov model
HMW	high(er) molecular weight
HMW-GS	high-molecular-weight glutenin subunits
HPLC	high-performance liquid chromatography

HPO ₃	phosphate
HPR	hydroxypyruvate reductase
HR	hypersensitive response
HRGP	hydroxyproline-rich glycoproteins
HRP	horseradish peroxidase
HSP	heat shock protein
H ₂ O ₂	hydrogen peroxide
H ₃ PO ₄	phosphoric acid
H ₂ S	hydrogen sulfide
<i>H. sapiens</i>	<i>Homo sapiens</i>
HSO ₃ [−]	sulfite
HSQC	hetero single quantum correlation
HUPO	human proteome organization
IAA	indole-3-acetic acid
IAC	inhibitor affinity chromatography
ICAT	isotope-coded affinity tag
ICL	isocitrate lyase
ICP	inductively coupled plasma
ICPL	isotope-coded protein label
ID	identification
IEF	isoelectric focusing
Ig	immunoglobulin
IgG	immunoglobulin G
IgE	immunoglobulin E
IGF	insulin-like growth factor
IM	inner membrane
IMAC	immobilized metal-ion affinity chromatography
IMS	intermembrane space
IPG	immobilized pH gradients
IPI	international protein index
IPTG	isopropyl-β-D-1-thiogalactopyranoside
IRGSP	international rice genome sequencing project
IRMPD	infrared multiphoton dissociation
IRRI	international rice research institute
ISG	interferon-stimulated gene
iTRAQ	isobaric tags for relative and absolute quantitation
j	junction
JA	jasmonic acid
JIP	jasmonate-induced protein
K	potassium
KBr	potassium bromide
KCBP	kinesin-like calmodulin-binding protein
KCl	potassium chloride
kDa	kilodalton
KH ₂ PO ₄	potassium phosphate

KIS	kinase interaction sequence
KNO ₃	potassium nitrate
<i>L. albus</i>	<i>Lupinus albus</i>
LAC	lectin affinity chromatography
LADH	(dihydro)lipoamide dehydrogenase
<i>L. bicolor</i>	<i>Laccaria bicolor</i>
LB	lysis buffer
LB-TT	lysis buffer containing thiourea and tris
LC	liquid chromatography
LCM	laser-capture microdissection
LC-MS	liquid chromatography-mass spectrometry
LC-MS/MS	liquid chromatography-tandem mass spectrometry
LDH	lactate dehydrogenase
LEA	late embryogenesis abundant
Leu	leucine
LHC	light harvesting complex
LiCl	lithium chloride
<i>L. japonicus</i>	<i>Lotus japonicus</i>
<i>L. maculans</i>	<i>Leptosphaeria maculans</i>
LMK	localization motif database
LMMs	lesion mimic mutants
LMW	low(er) molecular weight
LMW-GS	low-molecular-weight glutenin subunits
LOPIT	localization of organelle proteins by isotope tagging
LP	long pass
LRR	leucine-rich repeat
LSU	large subunit
LTP	lipid transfer protein
LTRE	low-temperature-responsive element
luc	luciferase
Lys	lysine
MA	matrix
MAC	mt-affinity chromatography
MAD	multiple-wavelength anomalous dispersion
MALDI	matrix-assisted laser desorption ionization
MALDI-TOF-MS	matrix-assisted laser desorption ionization time-of-flight mass spectrometry
MAP	myrosinase-associated protein
MAPs	microtubule-associated proteins
MAPK	mitogen-activated protein kinase
MAR	matrix attachment region
Mbp	mega base pair
MBP	myelin basic protein
MCAT	mass-coded abundance tagging
MCP	multichannel plate

MCF	mitochondrial carrier family
MDH	malate dehydrogenase
ME	2-mercaptoethanol/ β -mercaptoethanol
MES	2-(<i>N</i> -morpholino)ethanesulfonic acid
Met	methionine
MFP	multifunctional protein
<i>M. gracillis</i>	<i>Mammillaria gracillis</i>
<i>M. graminicola</i>	<i>Mycosphaerella graminicola</i>
<i>M. grisea</i>	<i>Magnaporthe grisea</i>
MgCl ₂	magnesium chloride
Mg	magnesium
<i>M. graminicola</i>	<i>Mycosphaerella graminicola</i>
MIAPe	minimum of information to describe a proteomics experiment
MKP	MAP kinase phosphatase
<i>M. loti</i>	<i>Mesorhizobium loti</i>
MMPs	matrix metalloproteinases
MMSDH	methylmalonate-semialdehyde dehydrogenase
Mn	manganese
Mo	molybdenum
MOAC	metal oxide/hydroxide affinity chromatography
Monocot	monocotyledonous
MOPS	3-[<i>N</i> -Morpholino]propanesulfonic acid
<i>M. pucida</i>	<i>Mimosa pucida</i>
MQ	milli Q
<i>M_r</i>	relative molecular mass
mRNA	messenger ribonucleic acid
<i>M. sativa</i>	<i>Medicago sativa</i>
MS	mass spectrometry/spectrometric
MS ⁿ	multiple mass scans
MS/MS	tandem mass spectrometry
MT	microtubule
<i>M. truncatula</i>	<i>Medicago truncatula</i>
MudPIT	multidimensional protein identification technology
MW	molecular weight
<i>M. xanthus</i>	<i>Myxococcus xanthus</i>
<i>m/z</i>	mass-to-charge
n	nano
N	nitrogen
Na ²⁺	sodium ion
NAC	nascent polypeptide-associated complex
NaCl	sodium chloride
Na ₂ CO ₃	sodium carbonate
<i>N. benthamiana</i>	<i>Nicotiana benthamiana</i>
NAD	nicotinamide adenine dinucleotide

NADP	nicotinamide adenine dinucleotide phosphate
NADPH	reduced form of nicotinamide adenine dinucleotide
NaF	sodium fluoride
NaHCO ₃	sodium bicarbonate
NaI	sodium iodide
NaOH	sodium hydroxide
NBS-LRR	nucleotide binding site-leucine rich repeat
NBT	nitro-blue tetrazolium chloride
<i>N. commun</i>	<i>Nostoc commun</i>
NCBI	national center for biotechnology information
ND	nondormant
NDPK	nucleoside diphosphate kinase
ND:YAG	neodymium-doped yttrium aluminium garnet
Ne	neon
NEDD	neural precursor cell-expressed developmentally down-regulated
NEPHGE	nonequilibrium pH gel electrophoresis
<i>N. haematococca</i>	<i>Nectria haematococca</i>
NH ₃	ammonia
NH ₄ Cl	ammonium chloride
NH ₄ HCO ₃	ammonium bicarbonate
(NH ₄) ₂ SO ₄	ammonium sulfate
Ni	nickel
NMR	nuclear magnetic resonance
NO	nitric oxide
Nod	nodulation
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
NOMT	naringenin 7- <i>O</i> -methyltransferase
NP-40	Nonidet P-40
NR	nitrate reductase
nrPTK	nonreceptor protein tyrosine kinase
nrPTP	nonreceptor protein tyrosine phosphatase
NSF	national science foundation
<i>N. tabacum</i>	<i>Nicotiana tabacum</i>
NTPP	amino-terminal propeptide
OASTL	<i>O</i> -acetylserine(thiol)lyase
OB	oil body/ies
1D	one-dimensional
ODE	ordinary differential equations
1-DGE	one-dimensional gel electrophoresis
1D LC-MS	one-dimensional liquid chromatography–mass spectrometry
1D LC-MS/MS	one-dimensional liquid chromatography–tandem mass spectrometry
OEC	oxygen evolving complex
OEE	oxygen evolving enhancer

OM	outer membrane
OPPP	oxidative pentose phosphate pathway
ORF	open reading frame
<i>O. sativa</i>	<i>Oryza sativa</i>
OsCIA	<i>Oryza sativa</i> cold-induced anther protein
O ₃	ozone
P	phosphate
<i>P. abies</i>	<i>Picea abies</i>
PAG	polyacrylamide gel
PAGE	polyacrylamide gel electrophoresis
PAL	phenylalanine-ammonia lyase
PAO	phenylarsine oxide
PAP	pokeweed antiviral protein
PAS	periodate/Schiff's base
PBZ	probenazole-inducible proteins
PC	principal component
PCA	principal component analysis
PCD	programmed cell death
PCR	polymerase chain reaction
Pb	lead
PDB	protein data bank
PDC	pyruvate dehydrogenase complex
PDH	pyruvate dehydrogenase
PDI	protein disulfide isomerase
PEB	protein extraction buffer
PEG	polyethyleneglycol
<i>P. euphratica</i>	<i>Populus euphratica</i>
<i>P. falciparum</i>	<i>Plasmodium falciparum</i>
PFCA	protein fragment complementation assays
PGIP	polygalacturonase inhibitor proteins
PGK	phosphoglycerate kinase
PGMA	phosphoglycerate mutase
PGR	plant growth regulator
pH	power of hydrogen
Phe	phenylalanine
PhIAT	phosphoprotein isotope-coded affinity tag
PhIST	phosphoprotein isotope-coded solid-phase tag
pI	isoelectric point
PIP	plasma membrane intrinsic protein
pK	negative logarithm of the dissociation constant K ($-\log K$)
PLC	phospholipase C
PLD	phospholipase D
PM	plasma membrane
<i>P. meyeri</i>	<i>Picea meyeri</i>
PMC	pollen mother cells

PMEI	pectin methylesterase inhibitors
PMF	peptide mass fingerprinting
PMIs	plant-microbe interactions
PMSF	phenyl methyl sulfonyl fluoride
PMT	photomultiplier tube
PNGase	peptide <i>N</i> -glycosidase
POMC	pro-opio-melanocortin
POR	protochlorophyllide oxidoreductase
PO ₃ [−]	phosphite
POX	peroxidase
<i>P. patens</i>	<i>Physcomitrella patens</i>
PPDB	plastid proteome database
PPI	protein–protein interaction
ppm	part per million
PP2A	protein phosphatase 2A
PR	pathogenesis-related
PRP	proline-rich proteins
Pro	proline
Pro-Q DPS	pro-Q diamond phosphoprotein stain
Pro-Q EPS	pro-Q emerald glycoprotein stain
PS	photosystem
PTC	phenylisothiocyanate
PTB	phosphotyrosine binding
pThr	phosphothreonine
PTK	protein tyrosine kinase
PTMs	post-translational modifications
PTPs	protein tyrosine phosphatases
PTSs	peroxisomal targeting signals
pSer	phosphoserine
<i>P. sativum</i>	<i>Pisum sativum</i>
<i>P. strobes</i>	<i>Pinus strobus</i>
PSD	post-source decay
PSI	position-specific iterative
PSI-MI	proteomics standards initiative–molecular interactions
<i>P. syringae</i> pv. <i>tomato</i>	<i>Pseudomonas syringae</i> pv. <i>tomato</i>
pTyr	phosphotyrosine
<i>P. trichocarpa</i>	<i>Populus trichocarpa</i>
<i>P. triticina</i>	<i>Puccinia triticina</i>
PVDF	polyvinyl difluoride
PVPP	polyvinyl pyrrolidone
PVS	peptide vinyl sulfones
PVX	potato virus X
Q	quadrupole
QAE	quaternary amino ethyl
QM	quantum mechanical

QSAR	quantitative structure activity relationship
QTL	quantitative trait loci
Q–TOF	quadrupole–time of flight
QqTOF	quadrupole–orthogonal time-of-flight
<i>R. communis</i>	<i>Ricinus communis</i>
<i>R. etli</i>	<i>Rhizobium etli</i>
RIO	right open reading frame
RIP	ribosome inactivating protein
RLK	receptor-like kinase
RMSD	root mean square deviation
RMPL	rice membrane protein library
RNA	ribonucleic acid
RNF	rhizobial nitrogen-fixing
RNP	ribonucleoprotein
ROC	receiver operating characteristic
ROS	reactive oxygen species
RP	reverse phase
RP-HPLC	reverse-phase-high-performance liquid chromatography
RP-LC	reverse-phase liquid chromatography
RPMA	reverse protein microarrays
RPTP	receptor-like protein tyrosine phosphatase
<i>R. solani</i>	<i>Rhizoctonia solani</i>
RT	reverse transcriptase
RTK	receptor protein tyrosine kinase
RuBisCO	ribulose-1,5-bisphosphate carboxylase/oxygenase
RYMV	rice yellow mottle virus
SA	salicylic acid
SAD	single-wavelength anomalous dispersion
<i>S. aegyptiaca</i>	<i>Suaeda aegyptiaca</i>
SalT	salt-induced protein
SAM	<i>S</i> -adenosyl-l-methionine
<i>S. arvensis</i>	<i>Sinapis arvensis</i>
SAR	scaffold attachment region
SAT	serine acetyltransferase
<i>S. aureus</i>	<i>Staphylococcus aureus</i>
SAGE	serial analysis of gene expression
SBML	systems biology markup language
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
SCFGs	stochastic context-free grammars
SCX	strong cation exchange
SDD	stomatal density and distribution
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SELDI	surface-enhanced laser desorption/ionization

SEMs	secondary electron multipliers
Ser	serine
SH	reduced sulfur
SHMT	serine hydroxymethyltransferase
SH2	src-homology 2
SIL	stable isotope labeling
SILAC	stable isotope labeling of amino acids in cell culture
SIPK	salicylic acid-induced protein kinase
<i>S. meliloti</i>	<i>Sinorhizobium meliloti</i>
smHSP	small heat shock protein
SMRPs	symbiosis-related proteins
SNARE	soluble <i>n</i> -ethylmaleimide-sensitive factor adaptator protein receptor
SNF	sucrose nonfermenting
<i>S. nodorum</i>	<i>Stagonospora nodorum</i>
SNP	sodium nitroprusside
SOD	superoxide dismutase
SO ₂	sulfur dioxide
SO ₃	sulfate
SPK	serine/threonine protein kinase
SPR	surface plasmon resonance
SPS	sucrose phosphate synthase
SRP	signal-recognition particle
SRPGS	SYPRO ruby protein gel stain
S–S	disulfide
<i>S. sclerotiorum</i>	<i>Sclerotinia sclerotiorum</i>
SSU	small subunit
STAT	signal transducer and activator of transcription
STY	ser/thr/tyr
SUBA	subcellular proteomic database
Suc	sucrose
SUMO	small ubiquitin-related modifier
SVM	support vector machine
SYMRK	symbiosis receptor kinase
TAC	tubulin affinity chromatography
TAAC	thylakoid ATP/ADP carrier
TAIR	the <i>Arabidopsis</i> information resource
TAP	tandem affinity purification
TAPa	alternative TAP
TAPi	improved TAP
TAT	twin-arginine translocation
TBP	tributyl phosphine
TBS	tris-buffered saline
TC	tentative consensus
TCA	trichloroacetic acid

TCAAEB	trichloroacetic acid/acetone extraction buffer
TCB	TEV cleavage buffer
TCTP	translationally controlled tumor protein
T-DNA	transfer deoxyribonucleic acid
TEM	transmission electron microscopy
TEMED	<i>n,n,n',n'</i> -tetramethylethylenediamine
TES	<i>n</i> -tris[hydroxy-methyl]methyl-2-aminoethanesulfonic acid
TEV	tobacco etch virus
TF	transcription factor
TFA	trifluoroacetic acid
TG	triglyceride
TGF β	transforming growth factor- β
THI	3-ketoacyl-CoA thiolase
Thr	threonine
TIGR	the institute for genomic research
TILLING	targeting-induced local lesions in genomes
TiO ₂	titanium dioxide
TLP	thaumatin-like protein
TM	transmembrane
TMD	transmembrane domain
TMT	tandem mass tags
TMV	tobacco mosaic virus
TMV-MP	tobacco mosaic virus-movement protein
TN	true negative
TOF	time of flight
TP	true positive
TPF	true positive fraction
TPI	triose phosphate isomerase
TPS	trehalose-6-phosphate-synthase
TPT	triosephosphate translocator
TPX	thioredoxin peroxidase
Tris	tris(hydroxymethyl)aminomethane
TROSY	transverse relaxation optimized spectroscopy
Try	tryptophan
TSV	total spot volume
Tyr	tyrosine
2D	two-dimensional
3D	three-dimensional
2-DE	two-dimensional electrophoresis
2-DGE	two-dimensional gel electrophoresis
2D-DIGE	two-dimensional difference in-gel electrophoresis
2D-LC	two-dimensional liquid chromatography
2D LC-MS/MS	two-dimensional liquid chromatography tandem mass spectrometry
2,4-D	2,4-dichlorophenoxyacetic acid

Ub	ubiquitin
UBDs	ubiquitin-binding domains
UBPs	ubiquitin-binding proteins
UCHs	Ub C-terminal hydrolases
UDP	uridine diphosphate
UFD	ubiquitin-fusion degradation
ULPs	ubiquitin-like proteins
<i>U. maydis</i>	<i>Ustilago maydis</i>
USPs	Ub-specific proteases
USSR	Union of Soviet Socialist Republics
UTR	untranslated region
UV	ultraviolet
V-ATPase	vacuolar H ⁺ -ATPase
2-VP	2-vinyl pyridine
vs	versus
VPE	vacuolar processing enzyme
VSR	vacuolar sorting receptor
VSV	valid spot volume
<i>V. unguiculata</i>	<i>Vigna unguiculata</i>
WAF	weeks after flowering
WAK	wall-associated kinase
WGA	wheat germ agglutinin A
Wx	waxy
Xe	xenon
XEG	xyloglucan <i>endo</i> -1,4- β - <i>d</i> -glucanases
XEGIP	xyloglucan-specific endoglucanase inhibitor protein
XML	extensible markup language
<i>Xoo</i>	<i>Xanthomonas oryza</i> <i>pv.</i> <i>oryzae</i>
YFP	yellow fluorescence protein
Y2H	yeast two-hybrid
ZE	zone electrophoresis
<i>Z. mays</i>	<i>Zea mays</i>
Zn	zinc
ZrO ₂	zirconium oxide

AN INTRODUCTION TO PROTEOMICS: APPLICATIONS TO PLANT BIOLOGY

Ralph A. Bradshaw

Although there is really no agreed starting date, the birth of proteomics, as a recognized discipline, is probably best marked by the introduction of two-dimensional (2D) gel electrophoresis (2-DGE) in 1975 [1, 2]. Perhaps somewhat surprisingly, this makes the field more than 30 years old and belies the fact that most people consider it to be quite a young area of activity. This can probably be attributed to the relatively slow progress in the early days and the rapid, essentially exponential growth that resulted from the determination of the sequence of the human genome [3], followed quickly by those of many more organisms, and the introduction and accelerated development of electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI)-based mass spectrometry (MS) [4]. Parallel development of immuno-based detection techniques, including notably the invention of monoclonal antibodies, molecular biological manipulations to allow the facile expression of native and engineered proteins, and improvements in the methodologies for protein isolation, also contributed greatly [5].

What was it then about these collected advances in technology that thrust proteomics into the forefront and distinguished it from the several decades of pioneering studies on the structure and function of proteins (usually referred to as “protein chemistry” [6]) that preceded it? Basically, it was the ability (perhaps even more accurately, the desire) to look at proteins not as individual components, which had been the root of the reductionist approach to biochemistry here to fore, but as complexes and interacting

partners with other proteins and cellular components. Indeed, it can be argued that the single most surprising finding that proteomic analyses have provided so far is the appreciation of the extent to which protein–protein interactions (PPIs) are manifested that has already fundamentally changed how we think about the organization and function of living cells. Novel methods, such as yeast two-hybrid (Y2H) analyses [7], were enormously valuable in opening up our appreciation of protein–protein networks, but it was MS, which allowed the unbiased identification of proteins in highly complex mixtures, that truly allowed the expansion of these determinations [8]. At the same time, MS also enabled the elucidation of post-translational modifications (PTMs) with a facility that previous methods could not remotely approach. Not only were old modifications, such as phosphorylations, detectable on an unprecedented scale but also new modifications (e.g., sulfation of serine and threonine [9]) were discovered with surprising regularity. The introduction of quantitative methods for MS, largely based on the incorporation of stable isotopes, was another milestone because analyses could be carried out allowing one to compare stimulated to unstimulated preparations both in terms of the appearance (or disappearance) of proteins but also with respect to changes in PTMs [4]. These measurements have had a major impact on signal transduction experiments and, when better correlated with RNA microarray data, will transform our understanding of cellular responses to external cues.

Clearly, proteomics has already provided substantial new understanding of biological systems, yet there are some who have been disappointed with the progress [10]. To appreciate both the successes and failures and to gain a proper feeling for what might be anticipated in the future, in terms of biology in general and plant sciences in particular, it is first necessary to grasp what proteomics encompasses and where the active areas of research are focused.

1.1 PROTEOMICS DEFINED

The genome and transcriptome of an organism represent easily defined, finite entities, albeit that in eukaryotes the latter can gain considerable complexity as compared to the former due to alternative splicing. The description of the complete proteome of the same entity is not presently imaginable, even for the simplest organism, and may well not be achievable; that is, it may be effectively infinite in size [10, 11]. The reasons for this are clear when one appreciates the scope of information required to provide this “complete description.” Table 1.1 provides a four-part definition for proteomics that outlines what this information encompasses [12]. Each component reflects different physical and chemical characteristics (and biological manifestations), and some have even achieved subdiscipline status (e.g., expression proteomics) [13]. A brief description of each part of the definition also provides the insight to understand the corresponding areas of current proteomics investigation.

The first component of the definition deals with the identification of all the proteins that make up the proteome. In prokaryotes, this is basically a one-to-one match-up of each protein with its gene. However, in eukaryotes, where multiple transcripts are

TABLE 1.1. The Components of Proteomics

Determination of the structure and function of the complete set of proteins produced by the genome of an organism, including co- and post-translational modifications
Determination of the all of the interactions of these proteins with small and large molecules of all types
Determination of the expression of these proteins as a function of time and physiological condition
The coordination of this information into a unified and consistent description of the organism at the cellular, organ, and whole-animal levels

Source: Adapted from reference 12.

produced from a single gene, it is already a much expanded number. Nonetheless, it is this aspect of proteomics that most closely parallels genomics and transcriptomics and has been a principal objective of large-scale analyses of complex mixtures, such as a protein extract of yeast [14]. There are actually two aspects here that should be distinguished: (1) the “missing” proteins that have not been identified in such large-scale analyses and (2) the “uncharacterized” proteins, revealed by genome analyses, but which as yet have not been defined in terms of structure or function. These two subsets, of course, overlap to some degree. However, matching up proteins to genes is only the beginning. To understand how a protein functions, one must also document the three-dimensional (3D) structure of each protein (challenging but certainly ultimately doable) and all covalent modifications that accompany the formation and maturation of that protein, beginning with the processing of the N-terminus of the nascent chain and ending with its ultimate destruction by proteolysis (usually at the hands of the proteasome). Not only has this not been achieved for any proteome, it likely will not be for a long time. New PTMs are continuously being found; and as technology improves, it is being discovered that determining these modifications is much more difficult than was first surmised. This is, in part, because earlier studies generally focused on proteins, which had one or two modifications, and because these were fully manifested on all molecules of the protein present. Partial occupancy was not a common occurrence for modifications associated with enzymatic requirements (activation or regulation). However, signal transduction appears to operate in a potentially different manner. In addition, background levels of what are likely spurious modifications, such as low-level phosphorylations that were not appreciated before MS achieved the present levels of sensitivity, add to the task of identifying those changes that are relevant (and, hence, part of the proteome description) *versus* those that are not.

The final part of the first component of the definition is the determination of function for each protein, and it is here that one must begin to seriously depart from the cataloging concept that is inherent in matching proteins to genes. While one can usually connect one or more functions (catalytic, recognitive, structural, etc.) to most proteins, ascribing single functions to proteins is at best a gross simplification of the situation. As already noted, PPIs that regulate activity in a variety of ways are widespread. The presence or absence of these interactions, along with the introduction

of many PTMs, are time-dependent; that is, they depend on the presence or absence of the participants or the causative agents, so that identifications alone are really inadequate to describe function in the cellular context.

The second part of the definition requires a detailed knowledge of all recognitive events, both stable and transient. Proteins clearly interact for a myriad of reasons [15]. Some are highly stable, and the complexes function as a unit tying together enzymatic activities that are a part of the same metabolic pathway. Others are more transient and assemble to permit an activity to occur and then disassemble to allow this activity to shut off. The elucidation of these interactions and the underlying networks that they describe, have been termed cell-mapping proteomics [13]. This, in fact, has been an active and quite productive area of proteomic research.

The third part of the definition clearly represents the greatest challenge in terms of experimental assessment because it introduces the element of time. To achieve valid data, it is necessary to know what proteins are present (or not) and what PTMs (both on and off) occur, generally as the result of some germane stimulus. It is often called “expression proteomics” [13]. Kinetic measurements are an inherent part of these determinations, and that greatly complicates experiments because almost all analyses have to be taken as specific time points. Hence, it is like trying to assemble a movie with only occasional snapshots. Signal transduction—and, to a considerable degree, expression proteomics—is closely related to what has been described as “systems biology” [4]. However, expression proteomics encompasses more than just signal transduction events, even if these are not readily separated from them. These include, for example, the regulation and function of metabolism, the cell cycle, apoptosis, and mitosis.

It is important to note that there is also a strong correlation between this type of proteomics experiment and mRNA microarrays. To date, the facility in generating such arrays and interrogating them has made them much more popular, but this is certainly about to change. As proteomic analyses become better, equivalent coverage of the proteome to what is represented by the microarray is now being achieved. Given the vastly superior information that the proteomic analyses provide (it is, after all, proteins that carry out most cellular functions, not RNAs), there will certainly be a shift in emphasis and data collection as expression proteomics develops. That said, it must be realized that the information to satisfy the third component of the definition will be the hardest to obtain and may never be really complete, at least in the foreseeable future.

The fourth part of the definition encompasses the bioinformatic area of proteomics. Bioinformatics is a poorly defined discipline and must really be considered in the context in which it is being used. It arose from the flood of data that first began appearing with genomic analyses and then rapidly spread to proteomics with the onset of large-scale, high-throughput experiments (from both MS and various protein arrays). Microarrays, of course, also contributed to this inundation. There has been a well-justified concern that software and data management would not be able to keep up with experimental data collection, and this has to some degree turned out to be the case. It is certainly clear that the analyses of many large-scale experiments have not exhaustively elucidated all the information that can be obtained from them. This

is, in part, why there has been a growing movement to ask authors to post raw data to germane websites so that other investigators with objectives distinct from those who generated the data can have the opportunity to analyze this data set without the expense of re-creating it. These efforts have been a part of several efforts to improve the quality of proteomics data and the accuracy of their reporting. It is important to remember that bioinformatics does cut across all of the “omics” fields, or perhaps more appropriately, ties all of them together so it is not an exclusive part of proteomics either. As noted previously, “this last part of the definition really also represents the ultimate goal of proteomics, if a field of study can be said to have a goal, which is to understand not only the nature of the components and how they change with time and condition but also how they integrate to produce a living entity” [12].

1.2 PROTEOMICS APPLIED

Practical proteomics—the application of experimental technologies to determining the information outlined in Table 1.1—is singularly dependent on only a few basic methods; however, these are manifested in a surprisingly broad number of variations. In the MS field alone, there is a great variety of instrumentation, software, search engines, and databases available; this has made it uniquely difficult to achieve common standards in both experimental design and data reporting. There are strengths and weaknesses with all of these; but some of the biggest problems lie with the wedding of biologists, with interesting questions and samples, and the analysts, with the skills and understanding of the technology required to make the germane measurements. The best interpretations of experimental data, in terms of what it means to the biology under study, are only as good as the quality of the data itself. Misidentification of proteins or the nature and location of PTMs is useless and is actually counterproductive in that it can lead to losses of time and wastes of research support [16]. Unfortunately, such problems are endemic in proteomics at present, and it will take a concerted effort by the scientific community to eradicate it.

One of the great attractions of proteomics is its broad applicability to any part of biology. Early and sustained focus on human and yeast, driven to some degree by the availability of genome data, were quite understandable. Human studies have devolved significantly into searches for biomarkers of disease (notably for cancer) (see, for example, reference 17), while yeast became the paradigm of choice for developing methods that could go deeper into any proteome and help deal with the dynamic range problem (the several orders of magnitude differences in protein concentration found in most biological samples) [14]. However, there was no scarcity of application of proteomic analyses to other areas, and plant biology was no exception. The availability of essentially complete genome sequences for *Arabidopsis*, rice, *M. truncatula* (barrel medic), and *P. trichocarpa* (black cottonwood), along with several additional sequencing projects in progress including corn, tomato, potato, and sorghum, have provided (and will provide) a rich database for plant proteomic analyses [18]. Indeed, studies are ongoing (see several chapters that follow) that have been devised to begin probing the proteomes of these and other plant systems.

When the term “proteome” was coined [10, 11], it was defined in terms of all the proteins expressed in a “genome, cell, or tissue.” Thus it has become quite popular to devise more proscribed projects that deal with readily obtained tissues. HUPO, for example, has sponsored plasma, liver, and brain proteome projects, among others, that are attempting to define the content of these materials in the hopes of finding biomarkers and drug targets for associated human pathologies [19]. This approach has also served plant proteomics well. Several of the chapters of this book deal with the specific proteomes of plant tissues such as seed, root, anther, and leaf and with important organelles such as chloroplasts and mitochondria (the latter, of course, not specific to plants). Since there are important unique functions associated with these entities, this subdivision of the whole proteome into more workable-sized targets should be a valuable approach.

It is also interesting, and important, to note that many of the proteomic analyses are quite similar to those already well advanced in mammalian cell paradigms. Cell signaling and the downstream PTMs induced, notably phosphorylation, is a well-traveled road in proteomics (there have been hundreds of papers on phosphoproteomics in the past few years); and although there are clearly unique substrates and pathways in plants, there is also much in common. At the same time, there are different signaling pathways in plants, and proteomic analyses offer the possibility to expand what is already known.

This volume illustrates all of these important points with contributions from leading laboratories describing the state of protein identifications, expression proteomics, and other unique analyses of interest in plant biology. On the whole, plant proteomics is not as extensively developed as other areas of concentration because it receives less general attention. This is regrettable from the point of view that technological gains in any type of proteomic experiment are almost always transferable, and it is a well-established principle that one biological system often holds clues to understanding processes in a parallel, if different, system. Thus, plant proteomics may well teach important lessons to those interested in mammalian paradigms (and vice versa). The scope of this collection provides ample opportunity for such “enrichment,” and all proteomic scientists would do well to peruse it.

In the same fashion that proteomics holds much promise for improving our understanding of cell, organ, and organism function in human and animal health, so also do proteomic analyses in plants. Improvements in yield and disease resistance as well as developing better species for meeting increasing requirements in human and animal nutrition will be the principal goals of these studies. This volume provides a broad perspective on the state of plant proteomics and details much of what this promise holds as well as the state of the field at present. It delineates exciting advances that are often less known or appreciated and which have already become an important part of the discipline of proteomics.

REFERENCES

1. Klose J. 1975. *Humangenetik* **26**:231–243.
2. O’Farrell PH. 1975. *J Biol Chem* **250**:4007–4021.

3. Ventner JC, et al. 2001. *Science* **291**:1304–1351.
4. Aebersold R and Mann M. 2003. *Nature* **422**:198–207.
5. Bradshaw RA and Burlingame AL. 2005. *IUBMB Life* **57**:267–272.
6. Lundblad RL. 2006. *The Evolution of Protein Chemistry to Proteomics: Basic Science to Clinical Application*, CRC Press, Boca Raton, FL, p. 295.
7. Fields S and Song O. 1989. *Nature* **340**:245–246.
8. Yates JR. 1998. *J Mass Spectrom* **33**:1–19.
9. Medzihradszky KF, et al. 2004. *Mol Cell Proteomics* **3**:429–445.
10. Huber LA. 2003. *Nat Rev Mol Cell Biol* **4**:74–80.
11. Cohen J. 2001. *Tech Rev* **104**:55–63.
12. Bradshaw RA. 2004. Proteomics today, proteomics tomorrow. In: *Biomedical and Pharmaceutical Applications of Proteomics* (Hondermarck H, ed.), Kluwer Press, Amsterdam, pp. 1–17.
13. Blackstock WP and Weir MP. 1999. *Trends Biotechnol* **17**:121–127.
14. Wolters DA, Washburn MP, and Yates JR 3rd. 2001. *Anal Chem* **73**:5683–5690.
15. Gavin AC, et al. 2006. *Nature* **440**:631–636.
16. Bradshaw RA, et al. 2006. *Mol Cell Proteomics* **5**:787–788.
17. Carr SA and Celis JE. 2006. *Mol Cell Proteomics* **5**:1719ff.
18. http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html#C_SEQ
19. <http://www.hupo.org/research/>

PART I

TECHNOLOGIES

GEL-BASED PROTEOMICS

Pier Giorgio Righetti, Paolo Antonoli,
Carolina Simò and Attilio Citterio

2.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

In the 1960s, two major events drastically boosted the performance of electrokinetic methodologies. The first one was the introduction of isoelectric focusing (IEF) via the theoretical work of Svensson [1, 2] and the synthesis of carrier ampholytes by Vestberg [3]. In IEF, proteins are captured along a particular point of their titration curve, where the local pH equals their isoelectric point (pI) value. By this method, proteins could be separated solely on the base of their surface charge. The second one was the discovery that polypeptide chains, when saturated with micelles of an anionic surfactant (notably sodium dodecyl sulfate, SDS), would reach a nearly constant charge to mass ratio, thus allowing their separation, in a sieving matrix (by polyacrylamide gel electrophoresis, PAGE), mostly according to their molecular mass [4]. The figure of about 1.4 g SDS bound per g of protein is often quoted as a typical stoichiometric value [5]. Both techniques represented a quantum jump over conventional zone electrophoresis (ZE), including disc-electrophoresis [6], in that it not only brought about a remarkable resolving power, but also permitted for the first time to assess some

important physicochemical parameters of polypeptide chains, such as their pI value (with a precision of about 0.02 pH unit) and their relative molecular mass (M_r , with a lower precision, on the order of ± 3 –5 kDa). ZE, on the contrary, in which contribution to migration is due to combination of protein charge, size, and shape, does not easily allow separating the different tributes. The only way of obtaining some discrimination is to resort to Ferguson plots [7], which, however, require a very cumbersome experimental setup. The third event, which brought us into the realm of modern proteomic analysis, was the report, in 1975, of 2D mapping, a procedure combining orthogonally first IEF and then SDS-PAGE in a large gel slab format. It was extraordinary that one of the very first papers reporting it, the most-quoted O'Farrell report [8], had elaborated the technique to its utmost sophistication, nullifying the attempts of the second and third wave of “discoverers” to get credit for it. He was able to resolve and detect about 1100 different proteins from lysed *E. coli* cells on a single 2D map and suggested that the maximum resolution capability might have been as high as 5000 different proteins. Apart from the meticulous attention to detail, major reasons for the advance in resolution obtained by O'Farrell, compared to earlier workers, included (1) the use of samples labeled with ^{14}C or ^{35}S to high specific activity and (2) the use of thin (0.8 mm) gel slabs for the second dimension, which could then be dried down easily for autoradiography. This detection method was able to reveal protein zones corresponding to one part in 10^7 of the sample (usually 1–20 μg was applied initially, since higher loads caused zone spreading, although up to 100 μg could be loaded). Coomassie Brilliant Blue (CBB), in comparison, was about three orders of magnitude less sensitive and could reveal only about 400 spots. For the first dimension, O'Farrell adopted gel rods of 13 cm in length and 2.5 mm in diameter. The idea was to run samples fully denatured, in what became known as the “O'Farrell lysis buffer (LB)” (9 M urea, 2% Nonidet P-40 (NP-40), 2% β -mercaptoethanol (2-ME) and 2% carrier ampholytes, in any desired pH interval). For the second SDS-PAGE dimension, O'Farrell used the discontinuous buffer system of Laemmli [9] and, for improved resolution, a concave exponential gradient of PAG (usually in the intervals 9–15 or 10–14%T, although wide porosity gradients (e.g., 5–22.5%T) were also suggested). It is thus seen that, since its very inception, O'Farrell carefully selected all the best conditions available at the time; it is no wonder that his system was adopted as such in the avalanche of reports that soon followed. It is no surprise that, with such a thorough methodological development, there were hardly any modifications to this technique, something that in the scientific world is a very rare event. We will now proceed to a more thorough description of these three gel methodologies that helped shape present-day proteomic analysis.

2.2 SDS-PAGE

SDS-PAGE, soon after its inception, became a most popular method for at least two scopes: for fast determination of M_r values of proteins and for assessing their purity. Today, its use covers additionally: monitoring protein integrity; comparison of protein composition of different samples; analysis of the number and size of polypeptide

subunits; Western blotting coupled to immuno-detection, and, of course, as a second dimension in 2D maps.

Conventional versus Discontinuous Buffers

Originally, the method exploited conventional buffer systems, typically a phosphate buffer around neutral pH, but soon discontinuous buffers were developed. The one that today is by far most popular is Laemmli's buffer [9], which exploits a glycine/HCl sweeping boundary in the stacking gel. Soon after, Neville [10] reported another type of discontinuous buffer, based on a boric acid/HCl sweeping boundary during the stacking phase. Although both systems seemed to be equivalent in terms of sharpening the zones, today the Neville system does not seem to have many followers. In both systems, though, the resolution of polypeptides below 14 kDa is not sufficient, since such smaller M_r chains co-migrate with the SDS front. One possibility is to use the technique of Schagger and von Jagow [11], which adopts an additional spacer gel between stacking and resolving gels, an increase of buffer molarity to 1 mol/L Tris-HCl, a lowering of gel pH to 8.4, and a change of terminating ion from Tris to Tricine. In this way the destacking of smaller M_r polypeptides and SDS front is much more efficient. The method claims linear resolution from as low as 1 kDa up to 100 kDa. However, this system too has problems: The running time is much longer, due to the high buffer molarity, and Tricine (as well as Bicine) is not compatible with ammoniacal silver stains. An improved variant has been described by Tastet et al. [12]: use of taurine as trailing ion, titrated to a pH value (8.05) coincident with the pK of Tris (the buffering counterion). This system ensures unstacking of smaller M_r chains, down to 6 kDa, while maintaining excellent resolving power up to 200 kDa. For the first time, 2D maps run in this system exhibited well-resolved chains in the M_r 6000–140,000 Da without resorting to porosity gradient gels.

Constant %T versus Porosity Gradients

When originally devised, SDS-PAGE was run in constant %T (T = total monomer concentration, expressed as the sum of acrylamide and cross-linker) gels, typically 8%T or 10%T. However, since the migration of SDS–protein complexes in a gel matrix is mostly determined by its size, it is not surprising that an improved fractionation can be achieved on gels of graded porosity (gels with increasing concentration gradients). In such gradients the migration velocity of a band is inversely related to time and varies exponentially with the distance traveled. Experimentally, it is found that not only does the rate of migration of components through a gel gradient vary inversely with time, but also that, after a sufficient time, a stable pattern will develop, in which the different components continue to move slowly but their relative positions remain constant. This has led to the concept of pore limit, defined as the distance migrated from the origin in a specific gradient after which further migration occurs at a slow rate directly proportional to time. Thus the technique of pore gradient electrophoresis is often referred to as pore-limit electrophoresis, although this latter term is perhaps an unfortunate one, since it has led to the common misconception that, once the pore

limit is reached, migration stops altogether. Since this is not so, it seems preferable to refer to the method as gel gradient electrophoresis.

For a complex mixture there seems to be little doubt that, under appropriate conditions, a gradient gel can give a resolution superior to that of a gel of single concentration, and it is not necessary for this that all components should have migrated as far as their pore limits. Part of this high resolution results from the fact that, throughout the run, the leading edge of any particular band is moving through more concentrated gel than the trailing edge and hence encounters greater resistance, resulting in a band-sharpening effect [13]. There are two further practical advantages of this method. First, since after the initial “sorting-out” process a relatively stable band pattern is formed, it is not necessary to control the electrophoretic conditions so precisely as in other electrokinetic procedures, at least for qualitative work. Second, once bands have migrated well into the gel and approached their pore limits, diffusion is greatly reduced, so that the gels can be kept unfixed for long periods of time with little loss of resolution [14]. One should notice that both systems (discontinuous buffers and porosity gradients) result in extra band sharpening. Originally, several authors adopted both of them, so as to obtain maximum resolution in 2D maps. However, today it is realized that the extra burden required for adopting both techniques does not pay off, so most scientists use either buffer discontinuities or gels of graded porosities, but not the combined treatment. In the case of porosity gradients, additionally, one should be aware that the best solution is to adopt linear gradients, not other types of gradients (e.g., convex, concave) since they are more difficult to be engineered and do not provide extra benefits. An interesting aspect of porosity gradients is the fact that they dramatically extend the linearity of the plot for estimation of M_r values. Whereas constant-porosity gels offer a limited linearity on given M_r windows (thus, for exploring large regions of M_r values several gels of different %T should be run) (Figure 2.1A), porosity gradient gels allow exploring a much wider range of M_r values while maintaining full linearity of the plot (Figure 2.1B) [15].

Peptide Mapping by SDS-Page

If a protein has already been extensively purified, it can be digested *in vitro* and then the peptides thus produced can be analyzed by SDS-PAGE. This is at the heart of the method of Cleveland et al. [16], which digests such purified proteins in a buffer already containing some SDS (0.5%), via a number of proteolytic enzymes, such as papain, chymotrypsin, or *S. aureus* protease, and then analyzes the peptides thus produced by SDS-PAGE. However, this method is of limited applicability, since in many cases purified protein samples are not available, yet the investigator requires information on the similarities between individual protein bands obtained by gel electrophoresis. In order to solve this problem, techniques for *in situ* peptide mapping have been developed and the methodology has been extensively described [17]. In these cases, the sample proteins are best separated by a preliminary electrophoretic step using virtually any of the usual methods (native electrophoresis, SDS-PAGE, IEF, and 2-DGE). The gels are then briefly stained and destained and placed on a transparent plastic sheet over a light box. Individual protein zones are then cut out with a scalpel or razor

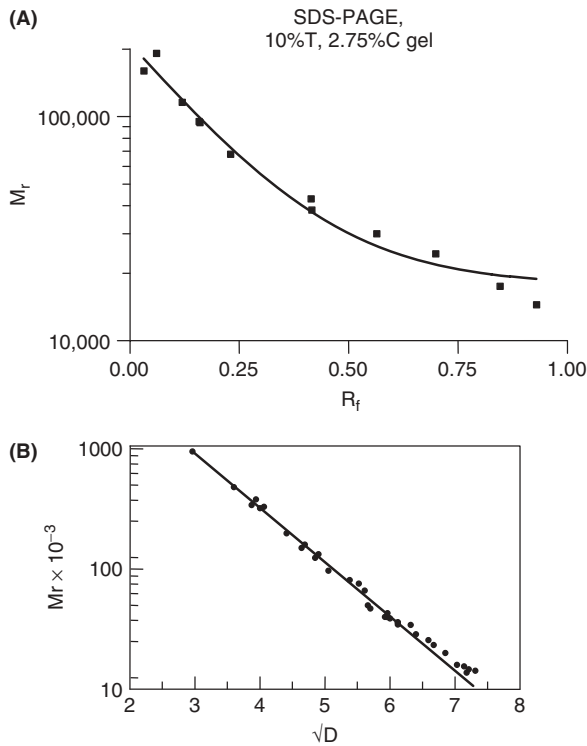


FIGURE 2.1. **A:** Calibration curve of $\log M_r$ versus R_f plotted with a series of markers in the 10,000–200,000-Da range, separated using a constant %T (10%T at 2.75°C) SDS-PAGE. Note the deviation from linearity. The markers are (M_r in kDa): myosin (194); RNA polymerase (β -subunit) (160); β -galactosidase (116); phosphorylase B (94); RNA polymerase (σ -subunit) (95); bovine serum albumin (68); ovalbumin (43); RNA polymerase (α -subunit) (38.4); carbonic anhydrase (30); trypsinogen (24.5); β -lactoglobulin (17.5); lysozyme (14.5). **B:** The method of Rothe and Maurer [15] for determining M_r using linear gradient SDS-PAGE. The plot shows the linear relationship between $\log M_r$ and \sqrt{D} (D is the distance traveled by the protein in the gel from the application point) for a series of 34 standard proteins covering the M_r 13,000–95,000 range, separated using a 3–30%T (at constant 8.4°C) linear gradient SDS-PAGE.

blade and trimmed to a size small enough to fit easily into the sample wells of an SDS-PAGE to be used for the peptide mapping stage. These gel pieces are overlaid with a protease solution, briefly incubated and then run by SDS-PAGE in a discontinuous buffer in presence of a porosity gradient (e.g., 10–25%T) so as to efficiently separate small-size peptides.

New Buffers and Systems

SDS-PAGE is a fundamental method for 2D map analysis, because as it represents the second dimension run and the run that will finally remain in the record, since it

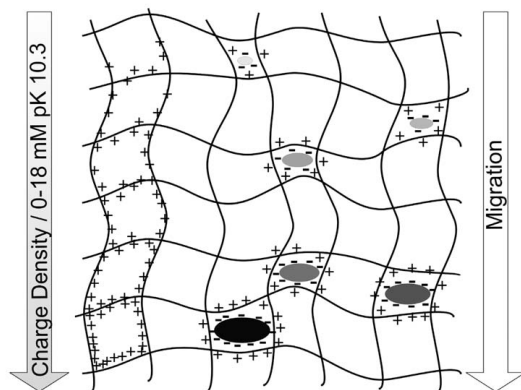


FIGURE 2.2. Pictogram of the “focusing” of protein–SDS complexes along a PAG containing a gradient of positive charges (in this case 0–18 mM Immobiline pK 10.3). Small proteins/peptides are focused in gel regions of low charge density (i.e., close to the application point), whereas larger proteins migrate progressively down the gel length toward regions of higher positive charge densities until reaching a “pI” due to balancing of positive gel charges with the negative charges of the various SDS–protein complexes.

is at the end of this step that proteins are stained, or blotted and extracted and further analyzed with the powerful tools today available in proteomics, such as MALDI time of flight (MALDI–TOF) MS of both (a) the intact polypeptide chain and (b) its peptides generated by proteolytic digestion. SDS-PAGE seems to have reached a plateau in terms of method development. It is hard to conceive that the technique could be further improved, since just about all possible methodological advances known in electrokinetic methods have been applied here. Yet, new developments appear to be in sight. Zilberstein et al. [18] have just reported a novel method, called “SDS-PAGE focusing,” that exploits a “steady-state” process by which the SDS–protein micelles are driven to stationary zones along the migration path against a gradient of positive charges affixed to the neutral polyacrylamide matrix (Figure 2.2). As the total negative surface charge of such complexes matches the surrounding charge density of the matrix, the SDS–protein complex stops migrating and remains stationary, as typical of steady-state separation techniques. As a result of this mechanism, the proteins are separated in an unorthodox way, with the smaller proteins/peptides staying closer to the application point and larger proteins migrating further down toward the anodic gel end. This results in a positive slope of the M_r versus migration plot, versus a negative slope in conventional SDS-PAGE. Particularly advantageous appears the ability of the present method to fine tune the separation of small size fragments and tryptic digests, where conventional SDS-PAGE usually fails. Additionally, by exploiting constant plateaus of charges, rather than gradients, it is possible to amplify the separation between species having closely spaced M_r values, down to a limit of ~ 150 Da. This increments the resolution by at least one order of magnitude as compared with standard

SDS-PAGE, where for a proper separation of two adjacent species, an M_r increment of ~ 3000 Da is needed.

2.3 IEF

The first dimension of 2D maps is an IEF step that can be performed either in conventional, soluble carrier ampholytes (CA-IEF) [1–3] or in immobilized pH gradients (IPGs) [19]. In this section we will describe briefly both methods and comment on their advantages and limitations. It is here emphasized that although in the vast majority of cases IPGs are today preferred over CA-IEF, at the light of some recent findings from our group also the latter methodology has substantial merits and is finding now interesting applications in proteome analysis, especially in its preparative version for sample prefractionation.

Conventional IEF in Soluble Amphoteric Buffers

CA-IEF was born as a preparative technique in a liquid phase, exploiting large-size columns filled with a sucrose density gradient as an anticonvective medium. When reports appeared on its use in PAG strips or thin tubes [20], its popularity grew exponentially until it became a household item in every lab. Although there are four major commercial brands available today, the formulas and general properties of only three of them are known: Servalyt, Ampholine, and Pharmalyte (Figure 2.3), with the Bio-Lyte product (from Bio-Rad) being undisclosed. It can be seen that Ampholine, still made according to the original patent by Vesterberg [3], comprises aliphatic oligoamino oligocarboxylic acids, obtained by reacting mixtures of oligoamines (3 to 9 nitrogen long) with unsaturated acids, such as acrylic and itaconic acids. Servalyt is made by first preparing oligoamines via reaction of ethylene imine with propylene diamine and collecting by distillation all products up to 400 Da in size. Such distilled products are then reacted with propane sultone and chloromethyl phosphonic acids for introducing the acidic counterions. Both commercial products are made so as to generate the most extended pH range (3–10), with narrow pH cuts (typically 2 or 3 pH units wide) being obtained via a preparative focusing process in large electrolyzers. Pharmalytes are made with an *ad hoc* synthesis, by reacting different amines and even amino acids with epichlorohydrin, so as to obtain directly narrow pH cuts (2–3 pH units wide). In the formula shown, it can be appreciated that mostly tertiary amines are obtained, with β -carboxyls generally as counterions (thus quite acidic, pK values in the pH range 2–3) and even peptide bonds (something that had been categorically excluded in Vesterberg's patent). Up to the present, not much has been known about the polydispersity, molecular mass distribution, and focusing properties of CAs, for any of the four commercial brands. Recently, our group has started an extensive investigation on such properties, by adopting a 2D technique consisting in a preparative focusing step in a Rotofor instrument, followed by analysis of the collected fractions by capillary ZE-MS (CZE-MS) [21–23]. The results have been impressive, as summarized here for the pH 4–6 interval, a very important pH range

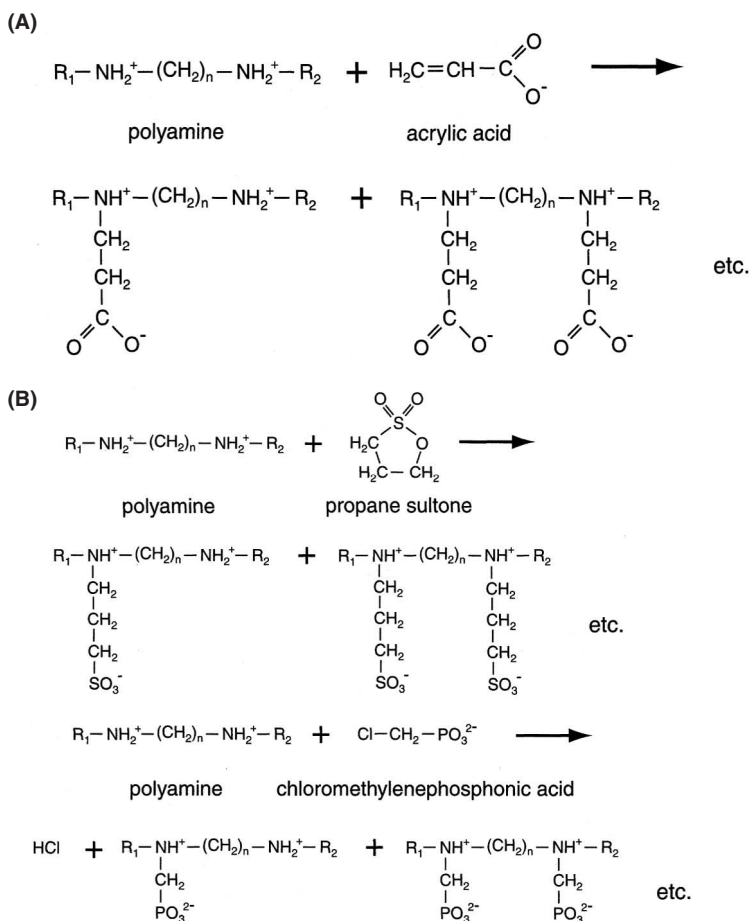


FIGURE 2.3. A: Chemical scheme for the synthesis of Ampholine. B: Chemical scheme for the synthesis of Servalyt. C: Hypothetical structure of a Pharmalyte constituent ampholyte containing six amines, one peptide bond, and three β -carboxylic groups. R corresponds to hydrophilic groups.

since it embraces the pI values of as many as 60% of all proteins expressed in living organisms. The findings: Ampholine contains 80 different M_r compounds, in the M_r interval 203 to 893 Da, for a total of 325 isoforms. Bio-Lyte consists of 66 different M_r species, in the M_r range 388 to 835 Da, for a total of 436 isoforms. Servalyt is made of 199 different M_r compounds, in the M_r interval 204 to 907 Da, for a total of 1302 isoforms. Pharmalyte pH 4–6.5 comprises 217 amphoteres, in the M_r range 150 to 1179 Da, for a total of 812 isoforms. Moreover, the vast majority of these species appear to focus as sharp zones, suggesting good buffering capacity and conductivity at pH=pI, in contrast with alkaline pH ranges (pH 8–10) where most species appear to focus poorly and thus exhibit rather shallow pH/mobility curves,

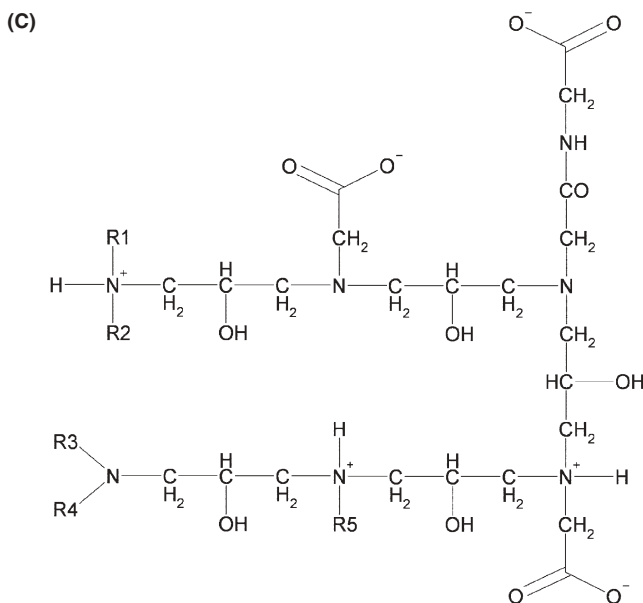


FIGURE 2.3. (Continued)

indicating poor buffering power and conductivity at $\text{pH} = \text{pI}$. This last observation is in agreement with findings from most scientists, who lamented severe difficulties in focusing basic proteins. At the light of these recent results, CA focusing in acidic pH ranges should be granted more credit for performances very close to those of IPGs. Additionally, although CA-IEF seems to have fewer followers in 2D map analysis, this technique seems to be quite valid in at least two important aspects of proteome analysis. In one instance, several groups have reported pre-fractionation of total cell lysates via preparative CA-IEF in a Rotofor instrument (or in a continuous free-flow electrophoresis device, such as the Octopus), enabling collection of up to 20 fractions to be further processed via 2D mapping and MS analysis (for reviews see references 23 and 24). In another experimental setup, CA-IEF is utilized, in a liquid vein, in a capillary format, followed by mobilization of the stack of focused proteins directly into and MS instrument [25]. Here too better recoveries of proteins and a higher success in protein identification are claimed over gel-based approaches.

IEF in IPG

It had to happen, it was in the air. CA-IEF, the raging technique of the 1960s and 1970s, had begun to show the crippling diseases of age. Here is the list of the major problems lamented by users: a medium of very low and unknown ionic strength; uneven buffering capacity and conductivity; unknown chemical environment; not amenable to pH gradient engineering; cathodic drift (plateau phenomenon) [26]. Some of these drawbacks were quite disturbing: for example, the fact that steady state, once

reached, would quickly begin to dissipate, with drift of the entire stack of focused CAs (and proteins) toward both anode and cathode, would lead to irreproducible spot position in the final 2D map. The same would occur due to variability in CA production from batch to batch (the synthetic process being, by definition, chaotic!). The risk of complex formation between some proteins and some CA chemical structures was also lamented [19]. When IPGs were first reported in 1982, it was apparent that all these problems had simply vanished [27].

IPGs are based on the principle that the pH gradient, which exists prior to the IEF run itself, is copolymerized, and thus insolubilized, within the fibers of a polyacrylamide matrix. This is achieved by using, as buffers, a set of six nonamphoteric weak acids and bases, having the following general chemical composition: $\text{CH}_2=\text{CH}-\text{CO}-\text{NH}-\text{R}$, where R denotes either two different weak carboxyl groups, with pKs 3.6 and 4.6, or four tertiary amino groups, with pKs 6.2, 7.0, 8.5, and 9.3 (available under the trade name Immobiline from Pharmacia-GE). A more extensive set, comprising 10 chemicals (a pK 3.1 acidic buffer, a pK 10.3 basic buffer and two strong titrants, a pK 1 acid, and a pK > 12 quaternary base), is available as “pI select” from Fluka AG, Buchs, Switzerland. During gel polymerization, these buffering species are efficiently incorporated into the gel (84–86% conversion efficiency at 50°C for 1 h). Immobiline-based pH gradients can be cast in the same way as conventional gradient PAGs, using a density gradient to stabilize the Immobiline concentration gradient, with the aid of a standard, two-vessel gradient mixer. As shown in their formulas, these buffers are no longer amphoteric, as in conventional IEF, but are bifunctional. At one end of the molecule is located the buffering (or titrant) group, and at the other end is an acrylic double bond, which disappears during immobilization of the buffer on the gel matrix. The three carboxyl Immobilines have rather small temperature coefficients (dpK/dT) in the 10–25°C range, due to their small standard heats of ionization (≈ 1 kcal/mol), and thus exhibit negligible pK variations in this temperature interval. On the other hand, the five basic Immobilines exhibit rather large ΔpK s (as much as $\Delta pK = 0.44$ for the pK 8.5 species) due to their larger heats of ionization (6–12 kcal/mol). Therefore, for reproducible runs and pH gradient calculations, all the experimental parameters have been fixed at 10°C.

Temperature is not the only variable that affects Immobiline pKs (and therefore the actual pH gradient generated). Additives in the gel that change the water structure (chaotropic agents, e.g., urea) or lower its dielectric constant, along with the ionic strength of the solution, alter their pK values. The largest changes, in fact, are due to the presence of urea: Acidic Immobilines increase their pK in 8 M urea by as much as 0.9 pH units, while the basic Immobilines increase their pK by only 0.45 pH unit. Detergents in the gel (2%) do not alter the Immobiline pK, suggesting that they are not incorporated into the surfactant micelle. For generating extended pH gradients, we use two additional chemicals that are strong titrants having pKs well outside the desired pH range. One is QAE (quaternary amino ethyl) acrylamide (pK > 12) and the other is AMPS (2-acrylamido-2-methyl propane sulfonic acid, pK 1.0). For the IPG run, the proteins are placed on a gel (or preferably adsorbed into the IPG strip during reswelling) with a preformed IPG. When the field is applied, only the sample molecules (and any ungrafted ions) migrate in the electric field. Upon termination of

electrophoresis, the proteins are separated into stationary, isoelectric zones. Due to the possibility of designing stable pH gradients at will, separations have been reported in only 0.1 pH unit-wide gradients over the entire gel length leading to an extremely high resolving power ($\Delta pI = 0.001$ pH unit, as opposed to a $\Delta pI = 0.01$ pH for CA-IEF).

It should not be taken for granted that IPGs took over and gained popularity just after their invention: At the beginning, we only knew how to make narrow pH intervals (only 1-pH-unit wide) and the technique was scorned and taken as an ancillary device in comparison with CA-IEF. It took several years of modeling in order to understand how to make extended gradients, first 2–3 pH units wide and then, as more chemicals and strong titrants were produced, up to the most extended pH intervals, covering the grounds from 2.5 up to pH 11. We here report, in Figure 2.4, one of the most popular IPG gradients adopted in proteome analysis: a nonlinear pH 4–10 gradient. It is the only “democratic” gradient ever devised, since it gives equal space (thus equal rights) to the variegated population of proteins in any proteome: Given the relative abundance of different species (shown in the histogram), it is clear that an optimally resolving pH gradient should have a gentler slope in the acidic portion and a steeper profile in the alkaline region [28].

Mixed-Type CA-IEF/IPG Gels

The fact that today, in gel-based proteome analysis, IPG strips have largely superseded CA-IEF does not mean that the latter technique has disappeared. On the contrary, today, in most cases, the first dimension run is performed in mixed-bed CA-IEF/IPG strips. There are several reasons for that: First of all, in not-properly washed IPG strips, ion boundaries would form and stop migrating somewhere between anode and

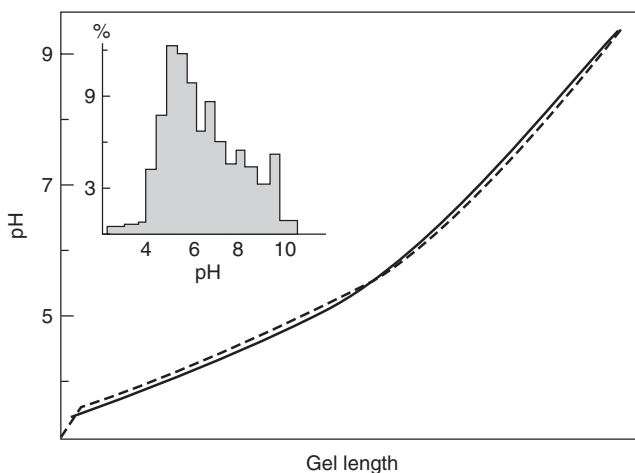


FIGURE 2.4. Nonlinear pH 4–10 gradient: “ideal” (solid line) and actual (---) courses. The shape of the “ideal” profile was computed from data on statistical distribution of protein pIs. The relevant histogram is redrawn in the figure inset.

cathode. These two boundaries would impede migration of applied sample toward its pI. Additionally, if one were to analyze native samples, often containing neutral buffers and salts (such a phosphate and NaCl), they would be quickly denatured upon application of the electric field. The reason: Two rather acidic and alkaline boundaries (containing, for example, free HCl and NaOH) would quickly form in the sample application drop and denature the proteins present in the sample, since such proteins are not any longer buffered as long as they are in the applied liquid droplet [29]. If the sample zone is added with CAs, they will act as buffering ions impeding such denaturation. Moreover, addition of CAs to sample and IPG strip often further increments sample solubility, since CAs can be adsorbed into the surfactant micelle and transform a neutral detergent into a kind of “zwitterionic” surfactant.

In conclusion of this section, a note of caution is here due. Although IPG strips are today the preferred media in 2D mapping, they are not trouble-free: Scientists lament that large proteins do not focus properly into IPG gels and are not transferred into the SDS-slab for the second dimension run. A partial remedy to it has been found by casting more porous IPG matrices, down to as low as 3.2%T (commercial strips consists of a 4%T polyacrylamide matrix) [30]. Additionally, one should note that the commercial IPG strips rarely meet the strict precision of the IPG technology. For instance, a strip claimed to cover a pH 3–10 interval would rarely reach in the alkaline region a pH value of 9.3 to 9.4. An IPG 4–7 strip would typically attain an upper pH value of 6.6. We have protested with the producers, but to no avail, so far. This is quite annoying, since it ruins the unique advantage of the IPG technology, namely the very high precision in pI measurements. On the good side, today it turns out that IPG strips (especially in the pH 4–6 range) can be used for separating peptides generated by tryptic digest of proteins, thus greatly contributing to their identification via MS analysis of the eluted peptides [31].

2.4 2D MAPS

For the last 30 years, 2-DGE has been the technique of choice for analyzing the protein composition of a given cell type and for monitoring changes in gene activity through the quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions. Proteins are usually the functional molecules and, therefore, the most likely components to reflect qualitative (expression of new proteins, PTMs) and quantitative (up- and down-regulation, coordinated expression) differences in gene expression. The 2D maps could be prepared by using virtually any combination of one-dimensional (1D) methods (not necessarily electrophoretic), but the one that has won universal recognition is that combining a charge (typically an IEF protocol) to a size (e.g., SDS-PAGE) fractionation, since this results in a more even distribution of components over the surface of the map. The 2D technique has been tremendously improved to generate 2D maps that are superior in terms of resolution and reproducibility. This new technique utilizes a unique first dimension that replaces the carrier ampholyte-generated pH gradients with IPG and replaces tube gels with gel strips supported by a plastic film backing. IPGs are so precise that they

allow excellent correlation between experimentally found and theoretically predicted pI values of both proteins and peptides. The field is so vast that it would take an entire book for a proper description. While we refer the readers to such a book [32] for a deeper insight, we will highlight here some important points worth considering for properly functioning 2D map analyses.

Sample Preparation

Proper sample preparation is the key to success in 2D map analysis. The sample has to meet the stringent requirements of the first dimension step, namely a focusing process either in conventional carrier ampholyte buffers or in IPG. The IEF step is quite sensitive to most substances present in the sample, such as high salt levels, presence of lipids, polysaccharides, mucins, and polyphenols, to name just a few. The lack of their removal results in severe loss of protein spots, vertical and horizontal streaking, and poor focusing. In most biological samples, whether of eukaryotic or prokaryotic origin, one of the ubiquitous interfering substances is nucleic acid, which in general leads to disturbances in the spot position in 2D maps and background streaking. Nucleic acids are visualized as horizontal streaks in the acidic part of the gel; furthermore, they can precipitate with the proteins when the sample has to be applied on the basic end of the IEF gel. Additionally, large-size DNA increments sample viscosity and can clog the gel pores, impeding protein entrance and causing also severe vertical streaking. An instructive example is offered in Figure 2.5: In bacterial extracts, lack of DNA removal results in massive vertical and horizontal streaking and fuzzy spots (panel A). When DNA is properly eliminated (panel B), sharp spots are observed all over the 2D plane. One can also appreciate a much larger number of spots present in the alkaline gel region, no doubt basic proteins formerly associated to DNA and thus lost in control samples still containing sizable amounts of DNA.

In plant extracts, the interfering substances (lignins, polyphenols, tannins, alkaloids, pigments) are so noxious that they have to be fully eliminated prior to 2D mapping. Their removal is usually achieved by precipitating proteins in 10% trichloroacetic acid (TCA) in cold acetone (-20°C). However, in recent reports, much better results are claimed if the sample is first solubilized in hot SDS and then subjected to TCA/acetone precipitation [33]. The first step seems to efficiently remove all interfering substances, while the second precipitation step is used to eliminate SDS that would be detrimental in the IEF step. Another point of major concern is accidental protein degradation during the solubilization step, prior to IEF analysis. The importance of safeguarding the sample against accidental proteolytic attack during preparation should never be underestimated. A case in point is the dramatic example of 2D maps of red blood cell membranes when processed in the absence or presence of proper protease inhibitor cocktails. Olivieri et al. [34] have shown that, in the absence of inhibitor cocktails, a massive protein degradation occurs, to the point at which, in the final 2D map, all high-molecular mass proteins (above 50 kDa) disappear from the map.

Another important aspect of sample preparation, ignored in the past, is the necessity of performing reduction and alkylation of $-\text{SH}$ groups prior to entering the electric field—that is, prior to the first dimension step. Failure to do that (or performing alkylation in between the first and second dimension runs) will result in spurious spots in the

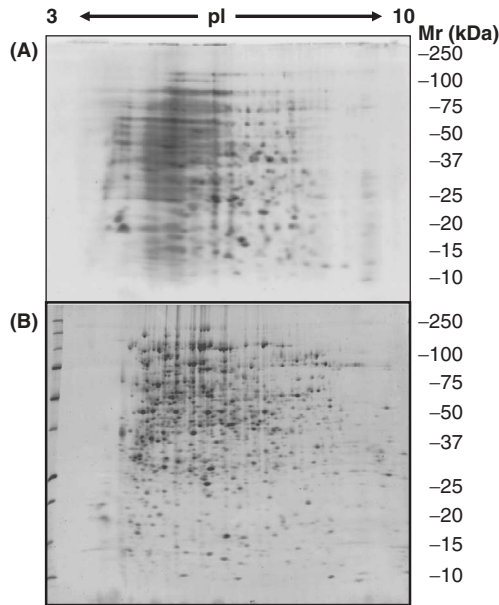


FIGURE 2.5. 2D maps of cytoplasmic proteins of selenite-reducing bacteria, under different conditions of DNA removal. **A:** Control map, untreated sample; **B:** After removal of DNA via two-phase partitioning (Antonioli and Righetti, unpublished data).

alkaline gel region (resulting from regeneration of $-S-S-$ bridges) due to formation of homo- and hetero-oligomers [35]. Additionally, severe streaking of alkaline spots ensues, in nonalkylated samples. The next question is how to perform the alkylation step. It is well known that the typical alkylating agent for cysteine residues, exploited in most protocols up to the present, is iodoacetamide. Indeed, our data have shown that alkylation with acrylamide and/or its derivatives are to be preferred, especially when dealing with 2D maps. The preferred solubilization cocktail for tissue and cell proteins, in use today, contains a mixture of 2 M thiourea and 7 M urea. We have demonstrated that thiourea is a scavenger of iodoacetamide, to the point at which all the amount added (typically of the order of 50 mM and higher) is consumed within a few min of incubation [36]. Thus, under these conditions, alkylation with iodoacetamide is incomplete and hardly quantitative. Although originally we had suggested alkylating with acrylamide, today we much prefer 2-vinyl pyridine (2-VP). First of all, due to the much higher reactivity of 2-VP, its concentration can be reduced to 20 mM. It should also be noted that the reaction here is conducted at neutral pH values (pH 7.0). This is optimal for the reactivity of both compounds (i.e., 2-VP and $-SH$ groups of Cys) since it is about halfway between the pK value of the $-SH$ groups (8.3) and the pK of the tertiary amino group of 2-VP (pK 5.4): under these conditions, the reaction is driven by the partial negative and positive charges, respectively, of the two reacting species and continues at a fast rate until full alkylation of all Cys residues. As

an additional bonus, this neutral pH during reaction ensures too its very high specificity, since at this pH value the ϵ -amino groups of lysine, being fully protonated, are unreactive. Thus, 2-VP appears to be among the best alkylating agents, since it couples 100% reactivity with 100% specificity, conditions rarely achievable with any other alkylating agent [37]. The reaction scheme is shown in Figure 2.6. There is an additional advantage in using 2-VP alkylation, especially when the protein sample is further analyzed by MS. In MS, all peptides labeled with 2-VP, due to the positive charge in this molecule, will give a much stronger signal, considering that most peptide analyses are performed in the positive ion mode. If this were not enough to convince 2D map aficionados to include an alkylating step in their protocols, one might want to consider another benefit—that is, the fact that 2-VP and alkylating agents blocking the free $-SH$ groups are also protective toward Cys residues. We have recently reported a noxious and unexpected artifact in proteome analysis: β -elimination (or desulfuration), which results in the loss of an H_2S group (34 Da) from Cys residues for proteins focusing in the alkaline pH region. Upon such an elimination event, a dehydro alanine residue is generated at the Cys site. In turn, the presence of a double bond in this position elicits lysis of the peptide bond, generating a number of peptides of fairly large size from an intact protein. The only remedy found to this noxious degradation pathway is the reduction and alkylation of all Cys residues prior to their exposure to the electric field. Alkylation appears to substantially reduce both β -elimination and the subsequent amido bond lysis [38].

Stains Compatible with MS Analysis

When doing quantitative proteomics, especially in separate gels and replicas, it is most important to use a stain that will be compatible with subsequent MS analysis and give a good linear response over as broad a range of concentrations as possible. Silver staining, of course, has been used extensively in the past, due to its extreme sensitivity

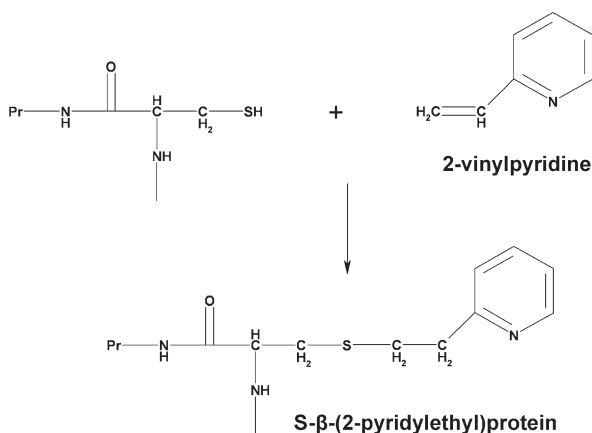


FIGURE 2.6. Alkylation process with 2-VP. This reaction produces S- β -pyridylethyl protein derivatives. Pr indicates the polypeptide chain.

(down to as little as 1 ng per 1-mm² spot). However, for differential proteome analysis, silvering is not a good procedure, first of all because it does not offer a good linear relationship and, secondly, because some of the reagents used (notably formaldehyde and glutaraldehyde) cross-link proteins and impede proper MS analysis. Micellar CBB has emerged as a stain of good reproducibility, good linearity, and decent sensitivity ($\sim 10\text{--}30$ ng per 1-mm² spot); however, the linearity would not seem to exceed two orders of magnitude in protein concentration range. Two new stains have emerged in the last few years, both of them fluorescent. The first one, reported in the year 2000, is SYPRO Ruby (Molecular Probes) [39]. SYPRO Ruby dye is a permanent stain comprised of ruthenium as part of an organic complex that interacts noncovalently with proteins. SYPRO Ruby protein gel stain (SRPGS) provides a sensitive, gentle, fluorescence-based method for detecting proteins in 1D SDS-PAGE and 2D maps. Proteins are fixed, stained from 3 h to overnight, and then rinsed in deionized water or dilute methanol/acetic acid solution for 30 min. The stain can be visualized using a wide range of excitation sources commonly adopted in image analysis systems including a 302-nm UV-B transilluminator, 473-nm second harmonic generation laser, 488-nm argon-ion laser, 532-nm yttrium–aluminum–garnet (YAG) laser, xenon arc lamp, blue fluorescent light bulb, or blue light-emitting diode. The sensitivity of SRPGS is superior to colloidal CBB stain or monobromobimane labeling and comparable with the highest sensitivity silver or zinc-imidazole staining procedures available. The linear dynamic range of SRPGS extends over three orders of magnitude, which is vastly superior to silver, zinc-imidazole, monobromobimane, and CBB stain [40]. The fluorescent stain does not contain superfluous chemicals (formaldehyde, glutaraldehyde, Tween-20) that frequently interfere with peptide identification in MS. While peptide mass profiles are severely altered in protein samples prelabeled with monobromobimane, successful identification of proteins by peptide mass profiling using MALDI-MS is easily performed after protein detection with SRPGS. Additionally, SYPRO Ruby stains glycoproteins, lipoproteins, calcium-binding proteins, fibrillar proteins and others, but will not stain extraneous nucleic acids.

The other stain that is taking momentum is Epicocconone, a novel fluorescent compound isolated from a fungus, sold under the trade name of Deep Purple [41]. Epicocconone, a heterocyclic organic fluorophore, spontaneously conjugates to Lys residues in proteins yielding a red fluorescent product. This change in fluorescence provides a novel approach for the sensitive quantification of proteins. Covalent binding of epicocconone to proteins is only stable at low pH (pH 2.5), and the fluor can be removed by washing at a higher pH. This pH-dependent reversible binding renders proteins amenable to subsequent analysis by MS and Edman-based sequencing. Deep Purple total protein stain (DPTPS) is a formulation of epicocconone used for staining electro-blots, SDS-PAGE, and IEF gels. DPTPS provides an order of magnitude greater sensitivity than SYPRO Ruby for gel and blot staining. Staining is in water, and procedures take between 3 h to 1 day for gels and < 30 min for blots [42]. As an extra bonus, DPTPS does not show the high background or speckling typical of other stains and is biodegradable, thus making disposal simple. Deep Purple staining does not affect enzymes activity or antibody binding. Epicocconone has also been

formulated into a simple-to-use, sensitive (ng/mL) rapid protein quantification kit (FluoroProfile) that can be used in solution or on polyvinylidene difluoride (PVDF) membranes. FluoroProfile is particularly suitable for situations where the protein sample is scarce or contains high concentrations of salts and detergents and for rapid determination of loading levels for 2D gels. Since Epicoconone binds to one of the primary cleavage sites of trypsin (Lys residue), this raises the possibility of incomplete cleavage and interference with MS analysis. However, a recent study has shown that this stain increases peptide recovery compared to SYPRO Ruby and can improve MS-based identification of lower-intensity protein spots [43].

Difference In-Gel Electrophoresis (DIGE) and Stable Isotope Codes for Quantitative Proteomics in 2D Maps

Up to recent times, differential quantitative proteomics in the 2D map format has been done by applying separately the controls and treated samples in separate gels (run in at least five replicas for proper statistical analysis) and then producing master maps and matching the spots via a number of software packages commercially available see reference [44] for review. Accurate quantitation relies upon computer analysis of a digitized representation of the stained gel. In the majority of cases, images are captured by laser densitometry, by phosphor imagery, or via a charge couple device (CCD) camera. Once the image has been digitized, a standard computer-assisted analysis of 2D gels includes at least the following three basic steps: (i) protein spot detection, (ii) spot quantitation, (iii) gel-to-gel matching of spot patterns. It is thus seen that this system is quite laborious and labor-intensive. In addition, the error introduced by the software in background cleaning and other spot recognition algorithms could induce additional variability to the data, further complicating spot acquisition from this type of approach [45]. The ideal situation, also in 2D map analysis, would be to have the ability to attach specific labels separately to the two populations under study, then mix the two samples (usually in a 1:1 ratio) and run them in a single gel. This would dramatically reduce variability and permit detection of protein modulation at much lower levels than those adopted when samples are run in separate gels (in this last case the threshold for accepting a meaningful variation is set at a factor of 2.0; that is, only spots whose quantity in gel B is at least twice that of the corresponding spot in gel A are accepted as significantly changed; thus we have 100% variation). Most of these shortcomings were eliminated by the introduction of a new methodology by Ünlü et al. [46], which they termed DIGE (GE-Healthcare). In this approach, two cyanine dyes (Cy₃ and Cy₅) are commonly used to label two different samples that are subsequently mixed and run on the same gel. These two dyes are mass- and charge-matched and possess distinct excitation and emission spectra. In later studies [47] a third dye (Cy₂) was used as an internal standard to provide a link for inter-gel comparison and to facilitate a more robust statistical analysis. These dyes undergo nucleophilic substitution with ϵ -amine groups of Lys residues present within the investigated sequences. It was also reported that the initial approach was to label all lysines through the use of excess dye, since incomplete reaction was expected to cause heterogeneity in electrophoretic mobility. However, it was found that carrying out

the labeling reaction to completion reduced protein solubility and resulted in protein precipitation before entering the gel, probably due to the replacement of primary amino groups with the hydrophobic cyanine dye. To avoid such an effect, the authors adopted what they termed a minimal labeling scheme in which only 1–2% of all the lysines were labeled.

As an alternative to Lys minimal labeling, Cy dyes (Cy3, Cy5) that target Cys rather than Lys residues in proteins prior to DIGE analysis were reported [48]. These “new” dyes were recently used for analysis of specific populations of cells in cancer tissues [49]. The idea behind this new set is the opposite of that on which the original DIGE is based. In other words, the minimal labeling of Lys residues is replaced by the maximum (saturation) labeling of the Cys residues. A close look at the data presented in both works suggests that the use of the new set of dyes will not add any advantage to the existing performance of this protocol; in fact, it might be detrimental. In the example given by Shaw et al. [48] the authors compared silver-stained 2D maps with their Cy3/Cy5 Lys and Cy3/Cy5 Cys counterparts. Such a comparison revealed superior sensitivity and a superior spot number in the silver-stained map. The inferior number of spots in the maps labeled with Cy dyes strongly indicates substantial precipitation of barely soluble proteins. Additionally, since one out of seven known proteins does not have any Cys residues, Cys labeling will automatically exclude such proteins from detection. In conclusion, we believe that “minimal” Lys labeling should be preferred to saturation Cys labeling [50].

2.5 CONCLUSIONS

In the brief space allotted to this chapter, we hope we have given the most important aspects of proteomic analysis via gel technologies. Here we will highlight some of the new findings. In the case of 1-DGE, perhaps the most interesting novel aspects come from the report by Zilberstein et al. [18], demonstrating SDS-PAGE under focusing conditions. In addition to the unique analytical advantages, there appear to be also important preparative aspects that can be carried through in a liquid phase. In terms of 1D CA-IEF/IPG, perhaps the important new findings come from the extensive work of Simò and co-workers [21–23] on the identification of the vast number of CA species, especially abundant in the acidic pH intervals. This structural analysis might open the way to the synthesis of novel commercial brands, endowed with better characteristics, especially in the basic pH ranges, where most carrier ampholyte buffers fail, and thus launch anew the use of CA-IEF. In terms of 2D maps, important aspects of sample pretreatment have been discussed, including removal of interfering substances in plant samples, DNA removal, and reduction and alkylation prior to any electrophoretic step. What will the future bring? It is hard to tell, but clearly gel techniques, especially in the analytical applications, will stay with us for a long time. Although liquid-phase methodologies are to be preferred, it is hard to envision abolishing a gel in a 2D map analysis, where the final large-size gel matrix indeed acts as a fraction collector. However, in their preparative version, more and more electrophoretic techniques will have to rely on liquid-phase approaches, since recoveries from a full liquid phase will

always be close to 100%. In fact, the Rotofor, the multicompartment electrolyzers with Immobiline membranes, and the free-flow IEF in the Octopus, to name just a few, are all liquid-based technologies.

2.6 FIVE-YEAR VIEWPOINT

It is hard to give a vision of the future, even immediate, considering that meteorological predictions are hardly reliable over periods of more than one week! Not to mention the most (in)famous five-year plans of Stalin in the now defunct USSR: a complete disaster. Rather than predicting, we will state here our wishes for improving the performance of 2D techniques, in the hope that our plea will reach those companies that produce the various chemicals, gels, and strips that we use daily in the lab. Here they are:

1. It would be highly desirable that within a few years all the IPG strips would be made more porous, moving down from the standard recipe (4%T) adopted by all companies to a more comfortable 3.5%T (or lower) fiber content. The increment of porosity will allow better entrance of large proteins and better transfer from the first- to the second-dimension run. Additionally, such “soft” matrices will overswell and allow loading more sample.
2. Even more urgent, there should be a re-formulation of all IPG strips, here too made by all companies. It is astonishing that what is offered today on the market falls short of expectations. No one of the pH ranges of the strips we routinely use meets the stringent requirements of respecting the pH ranges indicated. Just as an example, when one is running a pH 4–7 IPG strip, one should be aware that, although the starting point is quite close to the claimed value of pH 4, the end point does not go above pH 6.6–6.7. For the extended pH 3–10 IPG strips, the situation is even worse: Here too most commercial products start close to pH 3.0, but very few go above pH 9 [51]. This is true for all IPG strips offered in the market. This means that all the 2D maps so far published in the literature and posted at various websites are wrong by 0.2–0.3, up to as much as 1 pH unit on their pI axis. One could ignore the error in pI assessment, but surely one cannot brush aside the fact that a number of protein spots disappear from the map and are lost forever! Companies seem to be aware about that, and it is hoped that they will soon take action.
3. Although not mentioned in this chapter, there exists the problem of spot overlap in 2D map analysis. In a total cell/tissue lysate, also due to overcrowding in the pH 4–6 region (where > 50% of protein focus), rarely each individual stained spot will be a singlet (i.e., will comprise a single protein entity) but more likely will be a doublet or a triplet or even a higher envelope of species (in some occasions we have counted sextuplets) [52–55]. This aspect, unnoticed up to the present, has some severe implications in differential 2D map analysis: It could just happen that a protein, present in a multiplet, could be up- or down-regulated or be modified and move to another pI region in the 2D map

where other multiplets exist, thus being nested there. This would alter the data of differential 2D map processing and even obscure important changes. Pre-fractionation of a complex sample and running very narrow pI ranges (e.g., encompassing only 0.5 pH units, instead of the customary 2-pH-unit spans) will lessen the problem.

ACKNOWLEDGMENTS

This work was supported by a grant from the European Community, Contract No. 12793, project “Allergy Card,” by PRIN-2006 (MIUR, Rome) and by grants from Fondazione Cariplo (Milan, Italy). C. Simó is the winner of a Marie Curie fellowship from the European Community.

REFERENCES

1. Svensson H. 1961. *Acta Chem Scand* **15**:325–341.
2. Svensson H. 1962. *Acta Chem Scand* **16**:456–466.
3. Vesterberg O. 1969. *Acta Chem Scand* **23**:2653–2665.
4. Shapiro AL, Viñuela E, and Maizel JV. 1967. *Biochem Biophys Res Commun* **28**:815–820.
5. Pitt-Rivers R and Impiombato FS. 1968. *Biochem J* **109**:825–830.
6. Ornstein L. 1964. *Ann NY Acad Sci* **121**:321–349.
7. Ferguson KA. 1964. *Metabolism* **13**:985–1002.
8. O’Farrell PH. 1975. *J Biol Chem* **250**:4007–4021.
9. Laemmli UK. 1970. *Nature* **227**:680–681.
10. Neville DM, Jr. 1971. *J Biol Chem* **246**:6328–6334.
11. Schägger H and von Jagow G. 1987. *Anal Biochem* **166**:368–379.
12. Tastet C, et al. 2003. *Electrophoresis* **24**:1787–1794.
13. Margolis J and Kenrick KG. 1968. *Anal Biochem* **25**:347–362.
14. Margolis J and Kenrick KG. 1969. *Nature* **221**:1056–1057.
15. Rothe GM and Maurer WD. 1986. One-dimensional PAA-gel electrophoretic techniques to separate functional and denatured proteins. In: *Gel Electrophoresis of Proteins* (Dunn MJ, ed.), Wright, Bristol, pp. 37–140.
16. Cleveland DW, et al. 1977. *J Biol Chem* **252**:1102–1106.
17. Andrews AT. 1998. Peptide mapping. In: *Gel Electrophoresis of Proteins: A Practical Approach* (Hames BD, ed.), 3rd edition, Oxford University Press, Oxford, pp. 213–235.
18. Zilberstein G, et al. 2007. *Anal Chem* **79**:821–827.
19. Righetti PG. 1990. *Immobilized pH Gradients: Theory and Methodology*, Elsevier, Amsterdam.
20. Righetti PG. 1983. *Isoelectric Focusing: Theory, Methodology and Applications*, Elsevier, Amsterdam.
21. Sebastiano R, et al. 2006. *Electrophoresis* **27**:3919–3934.

22. Simó C, et al. 2006. *Electrophoresis* **27**:4849–4858.
23. Simó C, et al. 2007. *Electrophoresis* **28**:1488–1494.
24. Righetti PG, et al. 2003. *Proteomics* **3**:1397–1407.
25. Simpson DC and Smith RD. 2005. *Electrophoresis* **26**:1291–1305.
26. Righetti PG, et al. 2005. *Electrophoresis* **26**:297–319.
27. Bjellqvist B, et al. 1982. *J Biochem Biophys Methods* **6**:317–339.
28. Gianazza E, et al. 1985. *Electrophoresis* **6**:53–56.
29. Righetti PG, Chiari M, and Gelfi C. 1988. *Electrophoresis* **9**:65–73.
30. Candiano G, et al. 2002. *Electrophoresis* **23**:292–297.
31. Krijgsveld J, et al. 2006. *J Proteome Res* **5**:1721–1730.
32. Hamdan M and Righetti PG. 2005. *Proteomics Today: Protein Assessment and Biomarkers Using Mass Spectrometry, 2-D Electrophoresis and Microarray Technology*, Wiley-VCH, Hoboken, NJ.
33. Song J, et al. 2006. *Electrophoresis* **27**:3144–3151.
34. Olivieri E, Herbert B and Righetti PG. 2001. *Electrophoresis* **22**:560–565.
35. Herbert B, et al. 2001. *Electrophoresis* **22**:2046–2057.
36. Galvani M, et al. 2001. *Electrophoresis* **22**:2066–2074.
37. Righetti PG. 2006. *J Chromatogr B* **841**:14–22.
38. Herbert B, et al. 2003. *Proteomics* **3**:826–831.
39. Berggren K, et al. 2000. *Electrophoresis* **21**:2509–2521.
40. White IR, et al. 2004. *Electrophoresis* **25**:3048–3054.
41. Bell PJJ, Epicocconone KP. 2003. *J Am Chem Soc* **125**:9304–9305.
42. Coghlan DR, Mackintosh JA, and Karuso P. 2005. *Org Lett* **7**:2401–2404.
43. Tannu NS, et al. 2006. *Electrophoresis* **27**:3136–3143.
44. Marengo E, et al. 2005. *Proteomics* **5**:654–666.
45. Wheelock AM and Buckpitt AR. 2005. *Electrophoresis* **26**:4508–4520.
46. Ünlü M, Morgan ME, and Minden JS. 1997. *Electrophoresis* **18**:2071–2077.
47. Knowles M, et al. 2003. *Proteomics* **3**:1162–1171.
48. Shaw J, et al. 2003. *Proteomics* **3**:1181–1195.
49. Kondo T, et al. 2003. *Proteomics* **3**:1758–1766.
50. Righetti PG, et al. 2004. *J Chromatogr A* **1051**:3–17.
51. Righetti PG and Antonioli P. Manuscript in preparation.
52. Pietrogrande MC, et al. 2002. *Electrophoresis* **23**:283–291.
53. Pietrogrande MC, et al. 2003. *Electrophoresis* **24**:217–224.
54. Campostrini N, et al. 2005. *Proteomics* **5**:2385–2395.
55. Pietrogrande MC, et al. 2005. *Electrophoresis* **26**:2739–2748.

MASS SPECTROMETRY-BASED PROTEOMICS: IDENTIFYING PLANT PROTEINS

Eveline Bergmüller, Sacha Baginsky, and Wilhelm Gruissem

3.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

The principle of mass spectrometry (MS) originated more than 100 years ago, when J. J. Thomson noticed that the movement of ions is proportional to their mass-to-charge (m/z) ratio. Since then, MS has developed further and now represents an established method to analyze biomolecules. The analysis of large compounds (such as peptides and proteins) by MS, was only possible in the early 1980s, when fast atom bombardment (FAB) was developed [1]. However, the major breakthrough in characterization of peptides and proteins came with the introduction of two soft-ionization techniques: in 1988, MALDI by Karas and Hillenkamp [2]; and in 1989, ESI by Fenn et al. [3]. Thus, today MS represents the most powerful tool for protein analysis and is used to characterize the proteome of many organisms (Figure 3.1).

MS is an analytical tool that measures the m/z ratio of molecular ions characteristic for a given compound. It is used to discover the composition of a sample, to identify unknown compounds, and to determine the structure of a compound by its fragmentation pattern. Furthermore, modern MS techniques allow the identification of PTMs—for example, phosphorylation, methylation, acetylation, and ubiquitination.

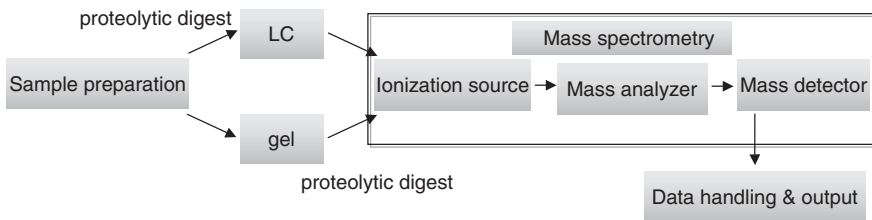


FIGURE 3.1. Schematic workflow of MS. Protein samples are proteolytic-digested and are separated either by SDS gel electrophoresis or by LC. The peptides/proteins are identified by MS. A mass spectrometer consists of three components: the ionization source, the mass analyzer, and the mass detector. Database search algorithms can interpret the obtained data.

MS can also be applied in quantitative proteome analysis. Here the proteomes of different organisms can be compared—for example, mutant and wild-type, stressed and nonstressed plants, and so on.

Today many MS instruments are available, having advantages but also limitations. High sensitivity, high mass accuracy, high mass resolution, and fast sample analyses are qualities of new state-of-the-art mass spectrometers. In this chapter we summarize the commonly used MS technologies used in plant protein/proteome analysis and highlight their application.

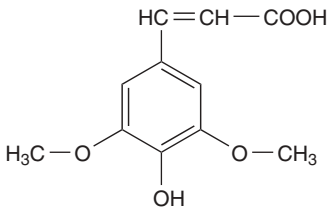
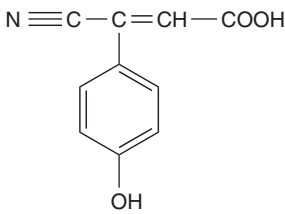
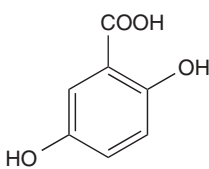
3.2 INSTRUMENTATION

A mass spectrometer is an instrument that measures the m/z ratio of ions. This can be achieved by ionization of the molecules, the separation of the different masses, and recording the intensities of the ions. Components of mass spectrometers are (a) the ion source that ionizes the molecule, (b) the mass analyzer, which separates the gas-phase ions according to their m/z values, and (c) the mass detector that records the signals. The data outcome of this process is either the mass of a molecule or ions derived from the parent ion after fragmentation when tandem MS (MS/MS) is applied. These fragmentation patterns are especially important to determine the amino acid sequence of a peptide and to identify sites of PTMs.

Ionization Source

The ionization source converts and transfers molecules into gas-phase ions, which are transported to the mass analyzer by a magnetic or an electric field. This is achieved by the loss/gain of an electron or the loss/gain of charge by protonation. Before the development of the soft-ionization techniques MALDI and ESI, electron ionization (EI) and chemical ionization (CI) were the most commonly used methods. However, these techniques were only appropriate for small molecules since they result in intensive thermal decomposition. Thus proteins and peptides could not be analyzed. The introduction of FAB in 1981 was a major step toward the analysis of proteins

TABLE 3.1. Summary of the Structures of Commonly Used Matrix Compounds for MALDI

Matrix	Structure
3,5-Dimethoxy-4-hydroxycinnamic acid (sinapinic acid)	
α -Cyano-4-hydroxycinnamic acid	
2,5-Dihydroxybenzoic acid	

[1]. The development of MALDI and ESI in 1988 and 1989 made it possible to characterize proteins by MS because no excessive fragmentation of large compounds occurred [2, 3].

3.3 MALDI

The principle of MALDI ionization relies on the co-crystallization of the analyte and the matrix with an excess of the matrix. The bombardment of the matrix/analyte with N_2 -laser results in an energy transfer and thus in the protonation of matrix and analyte. Ions are then detached from the crystal in a nonfragmentation manner that makes it possible to acquire the molecular mass of the compounds, and they are subsequently transferred into the gas phase. A high electric field (10–30 kV) accelerates the ions toward the mass analyzer.

Usually the matrix is an aromatic molecule, which absorbs and transfers energy at specific laser wavelength (337 nm). Table 3.1 gives a summary of the structures of commonly used matrix compounds for MALDI. Although the ionization mechanism

is still not fully understood, it is known that the nature of the matrix plays a role in the initial ionization of the compound. Usually the compounds are protonated, which leads to singly charged molecules $[M+H]^+$ that can be separated according to their m/z . Different mass analyzers can analyze the vaporized ions. Typically, MALDI is coupled to a TOF mass analyzer (MALDI-TOF-MS). One problem of this analysis is the low resolution due to different desorption and ionization times of molecules with the same molecular mass. This problem has been solved through the introduction of delayed ion extraction and reflectron technology. The delayed ion extraction (DE) results from an electrical field that is applied with a time delay to the laser pulse. During this delay, ions are allowed to spread in the source and higher-energy ions will move further away than lower energy ions with the same mass. Modern MALDI instruments are used in combination with a TOF-TOF analyzer, capable of MS/MS.

3.4 ESI

For the development of ESI, John Fenn was awarded the Nobel Prize in 2002 [3]. A modification of ESI is nano-ESI (nESI), where only a few microliters of sample are needed. A high voltage is applied to the liquid sample as it passes through a small capillary. Thus charged droplets are generated and a constant spray of 20–50 nL/min is formed. The evaporation of the solvent at the tip of the column results in gas-phase ions, which are directed under increasing vacuum to the mass analyzer. To generate gas-phase ions, solvents such as acetonitrile or 2-propanol are optimal. The charged molecules can be singly or multiply charged depending on the sample composition, the pH of the solvent, and the chemical nature of the peptides analyzed. Peptides smaller than 2000 Da are mostly doubly charged and sometimes singly charged. Peptides larger than 2000 Da often result in multiple charges. Multiple charges produce lower m/z ratio, and therefore large peptides give values in a range that can be measured.

3.5 MASS ANALYZERS

The mass analyzer separates peptides according to their m/z ratio. Different types of mass analyzers exist, all of them having advantages and disadvantages. In most cases, more than one mass analyzer is used for the analysis of peptides and thus results in MS/MS, a technique that provides information about the amino acid sequence of a peptide. Here a brief overview on common mass analyzers used in proteomics is given.

Quadrupole (Q)

A quadrupole consists of four parallel rods. Each two of the rods form an electrically connected pair, and a radio-frequency voltage is applied between one pair, thus generating an oscillating field. Ions with a specific m/z ratio are stabilized between the electrodes at a given ratio of voltage, while the majority collides with the electrodes.

Therefore, quadrupole has the characteristics of a mass filter, which allows the selection of particular ions. Quadrupoles are commonly used as triple quadrupoles (triple quad/Q) or in combination with a TOF mass analyzer, for MS/MS. A triple quad is the linear arrangement of three quadrupoles. Q1 and Q3 are mass filters, and Q2 represents a collision cell [4]. In Q1, ions are selected and subsequently fragmented in Q2 by the use of helium gas. Afterwards, the obtained fragments are analyzed by Q3.

TOF

This mass analyzer measures the flight time of ions until they reach the detector. An electrical field accelerates ions, and the time they take to reach the detector is measured. Ions with the same initial translational energy and m/z ratio have the same flight time, whereas heavier particles have a lower travel time. However, the mass resolution of a TOF analyzer is reduced due to the spreading of the initial energy. To increase the mass resolution of a TOF analyzer, a reflectron at the end of the tube was implemented [2]. The reflectron focuses ions with the same m/z ratio but different kinetic energies in space and time, though they arrive at the same time at the detector. Initially, the TOF analyzer was coupled to MALDI. However, it has now been successfully coupled to ESI. ESI and MALDI are often used in combination with Q-TOF analyzers. These instruments have the advantage of high resolution of the TOF analyzer and the efficient ion selection of the quadrupole [5, 6].

Ion Trap

The ion trap is a tandem-in-time mass analyzer, and it is functionally similar to a quadrupole analyzer [7, 8]. An ion trap consists of one ring electrode and two end-cap electrodes, in which the ions are trapped 3D. Both end-cap electrodes have small holes through which the ions enter and leave the trap. In an ion trap the ions are injected into the trap by pulsing of the ion gate, whereas in a quad analyzer the ions enter continuously. Compared to a triple quad, where selection, collision, and mass analysis occurs sequentially in three successively aligned devices, an ion trap performs all three steps in one device. One limitation of an ion trap is the loss of mass accuracy and resolution because of the so-called space-charge effect. Too many ions in the trap cause a distortion of the electrical field. Therefore, the settings have to be chosen, to obtain maximal signal and minimal space-charge effect.

For MS/MS, precursor ions can be selected by ejecting all ions with higher and lower m/z ratio. The stable trapping of a specific ion depends on the mass and charge, the size of the ion trap, and the oscillating frequency of the fundamental R_f , the potential applied the ring electrode. Collision-induced dissociation (CID) of the precursor ion with helium results in fragmentation ions that can be scanned and generate MS/MS spectra. Ion traps can produce several MS scans (MS^n) since they are acquired in time and not in space. The advantages of an ion trap are the low costs, the fast scanning rate, and suitable to high throughput.

Fourier Transform–Ion Cyclotron Resonance (FTICR)

In FTICR–MS, ions are trapped in a 3D space [9]. FTICR uses a static magnetic field in which ions are forced to rotate in an orbit, depending on their m/z values. In order to obtain a signal, ions need to be excited by applying an oscillating electrical field. If the frequency of the applied field is the same as the cyclotron frequency of the ions, the ions absorb energy, thus increasing their velocity (and the orbital radius) but keeping a constant cyclotron frequency. The changes of the radius of the circular motion results in an image current that can be measured at the detector plates of the mass spectrometer. Since the ions are not hitting the detector plates, it is a nondestructive analysis, whereas an electrical signal is measured over the time. Each ion is counted more than one time, resulting in a very high sensitivity, high resolution, and mass accuracy. The measured signals are Fourier-transformed to obtain a mass spectrum. The operation of FTICR instruments is cost-intensive, since they are technically very challenging. Nevertheless, these instruments are a very attractive and important analysis tool, due to their high mass accuracy, high resolution, and high sensitivity.

Orbitrap

The Orbitrap is the most recent mass analyzer that has been introduced to the market. It represents a highly modified ion trap with features of an FTICR. In an Orbitrap a constant electrostatic field is used for ion trapping, instead of an oscillating electric field, like in the ion trap. The electrostatic quadro-logarithmic field is created between a central spindle electrode and a coaxial outer electrode that resembles a barrel. The trapped ions orbit around the central spindle electrode and simultaneously oscillate in the axial direction. This oscillation generates an image current that can be measured by detector plates. The measured ion frequencies depend on the m/z ratio of the particle and are subsequently translated into a mass spectrum by Fourier transformation [10, 11]. This mass analyzer can be coupled to MALDI and ESI ionization sources. Orbitraps have advantages similar to those of FTICR, such as high resolution, high mass accuracy, high sensitivity, and a good dynamic range, and they have the advantage that they are less cost-intensive.

3.6 ION DETECTORS

Except for FTICR and Orbitraps, where the ion detection is nondestructive, mass analyzers have a destructive ion detector as a component of a mass spectrometer. The detector records the charge or the current that is induced when an ion hits its surface. The magnitude of current produced at the detector as a function of time is used to calculate the m/z and the intensity of the signal. Usually the number of ions is low and needs to be multiplied to obtain a significant signal. The intensity of peaks in a mass spectrum can be normalized and given as relative abundance in percentage. The most abundant peak is called the base peak and is normalized to 100. Most common detectors are secondary electron multipliers (SEMs) and microchannel plate (MCP) detectors.

3.7 SAMPLE PREPARATION

Sample preparation is the critical step in all proteomics research areas. For successful proteomics studies it is important to reduce the complexity of a sample’s proteome. Compared to other organisms, plants have additional characteristics that complicate protein analysis. They are rich in proteases and also in metabolites such as cell wall polyphenols and polysaccharides, starch, and lipids that interfere with the analysis [12]. Furthermore, they contain chloroplasts, which are highly abundant in proteins involved in photosynthesis [e.g., ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)] that prevent the detection of proteins with lower concentration. Since proteins are highly diverse, the extraction protocol has to be adapted for each group of proteins to be analyzed. To enhance the coverage of certain proteins, strategies for the fractionation of the proteome have been developed: serial extraction of proteins according their solubility, chromatographic fractionation, or purification of organelles (plastids, nuclei, mitochondria, vacuoles, etc; reviewed in references 12 and 13). Precipitation of proteins is helpful to concentrate proteins and to remove salts and other contaminants. TCA/acetone or phenol extraction methods are common precipitation procedures; however, they result in protein pellets that are difficult to resolubilize. For MS, the proteins should be concentrated; and in solution and high concentrations of phosphate, Tris and salts have to be avoided. Furthermore, detergents or solvents such as 3-[(3-choramidopropyl)dimethylamino]propane-1-sulfonate (CHAPS), Brij 35, Triton X-100, dimethyl sulfoxide (DMSO), glycerol, and Tween are not compatible with MALDI and ESI analysis. Mostly prior to analysis the proteins are digested with specific endopeptidases, and the resulting peptides are injected into the mass spectrometer. The choice of the endopeptidase depends on the protein sequence. For complete protein coverage, digestion with more than one peptidase is recommended. A very commonly used peptidase is trypsin, which cleaves C-terminal of Lys and arginine residues. Lys and Arg are positively charged amino acids, and therefore tryptic digestion of proteins results in doubly charged peptides, one positive charge at the N-terminus and one at the Arg/Lys residue. Table 3.2 gives a summary of some endopeptidases used in proteomics.

TABLE 3.2. Endopeptidases Commonly Used in MS and Their Specificity

Endopeptidase	Type	Specificity
Trypsin	Serine	K, R
Chemotrypsin	Serine	Y, F, W
ArgC	Cysteine	R
GluC	Serine	E
LysC	Serine	K
AspN	Metallo	D

3.8 PROTEIN IDENTIFICATION

The identification of proteins with MS can be done in a database-dependent or -independent manner. Peptide mass fingerprint (PMF) is one of the fastest methods to identify proteins, and it does not require MS/MS. Peptide fragmentation identification and *de novo* sequencing both need MS/MS data, while the latter one does not use a database as a reference. In the next subsection, peptide fragmentation identification and *de novo* sequencing are described as tools for protein identification.

Peptide Mass Fingerprint (PMF)

A mass spectrum of a peptide mixture that results from the proteolytic digest of a protein of interest can provide a PMF, which is specific for a given protein and thus allows its identification [14, 15]. A general approach for PMF protein identification from complex protein samples is the separation of proteins by 2-DGE, the excision of proteins from the gel, and subsequent enzymatic digestion (e.g., by trypsin). Afterwards the peptide mixture is analyzed by MALDI–TOF analysis and the MS spectrum is compared *in silico* with the genome of an organism (Figure 3.2). A computer program uses protein sequences and theoretically digests them with the respective enzyme. The masses of the originated peptides of each protein are calculated and compared to the protein of interest. Statistics are used to calculate the best match, which results in a respective score. PMF analysis is very applicable for high throughput analysis, since it is very fast. However, it has several limitations. The occurrence of more than one protein in one gel spot, the presence of contaminants or PTMs may complicate PMF identification. For the analysis of PTMs, MS/MS analysis is required.

Peptide Fragmentation Identification

Peptide fragmentation identification requires MS/MS and allows the unambiguous identification of proteins, even from complex protein mixtures. Thus, this analysis of peptides requires the fragmentation of peptides into different ions by CID. The pattern of the fragment ions can be interpreted by database searches, since the fragmentation spectrum of a given peptide is somehow predictable (Figure 3.2). Databases predict that the fragment ions of all available peptides in the database and the theoretical and measured, ion pattern are compared to identify matches. Thus, not only the peptide mass is measured, but also amino acid sequence information is provided. The correct peptide interpretation is based on MS/MS search algorithms that are provided by search engines such as Mascot™ [16], SEQUEST™ [17], and Phenyx™ [18, 19].

To obtain MS/MS spectra, two mass analyzers and a collision cell need to be coupled. The first mass analyzer measures the parent mass of the ion and the second records the fragment ions after CID. Fragmentation of peptides results in specific internal ions. The most abundant ions are b-ions, if the N-terminus is included; and y-ions, if the C-terminus is included [20, 21]. By calculating the mass difference between the ion series, amino acid sequence information is obtained. Besides these ions, other internal fragment ions are also obtained. The fragmentation pattern of a certain peptide is highly dependent on its amino acid sequence.

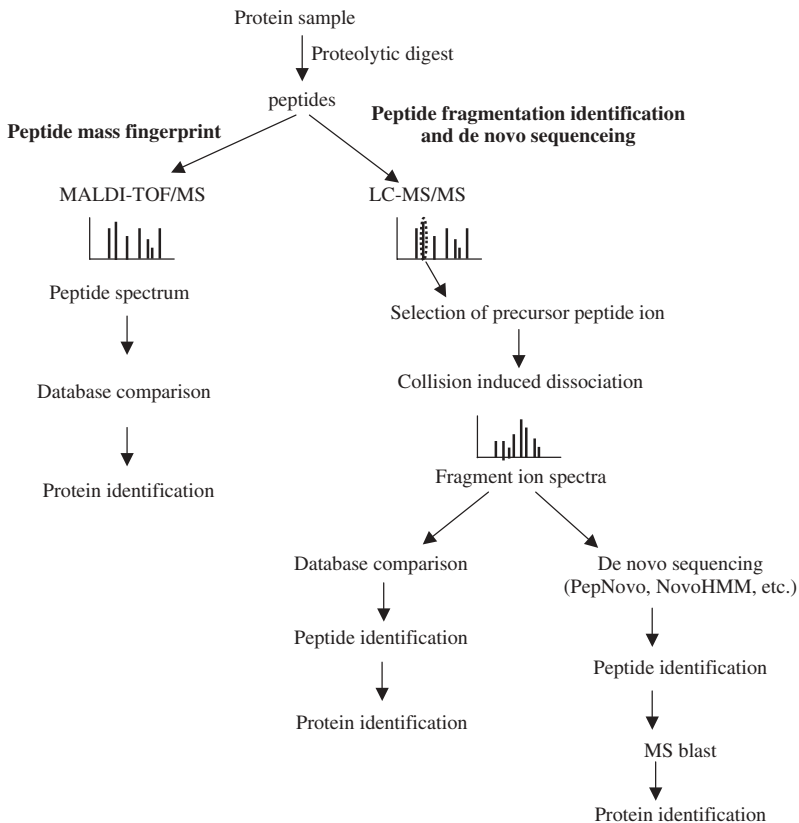


FIGURE 3.2. Differences between PMF, peptide fragmentation identification, and *de novo* sequencing. Proteins can be identified by PMF with MALDI–TOF/MS or by peptide fragmentation identification and *de novo* sequencing using ESI–MS/MS. Peptide fragmentation identification and *de novo* sequencing not only lead to the protein identification but also give information on the amino acid sequence.

ESI and MALDI can both be coupled to MS/MS. For example, ESI–Q–TOF and MALDI–TOF/TOF are instruments that give MS/MS spectra with very good mass accuracy and allow observation of ions in the low-molecular-mass range. This has the advantage that immonium ions of amino acids and diagnostic marker ions are obtained. Ion traps, which are commonly used for high-throughput proteomics, can perform MSⁿ, but they are limited to a lower mass range, lower mass accuracy, and lower resolution. Orbitrap and FTICR instruments can also be used for MS/MS and give spectra of very high resolution and very high mass accuracy. FTICR–MS is now often used for “top-down” protein characterization. Hence, intact proteins are sprayed into the mass analyzer and fragmented. Due to the high mass accuracy of the instrument, the resulting fragment ions can be interpreted. High mass accuracy of the precursor ion and good MS/MS spectra are especially important for the identification of PTMs

and also to allow the distinction of amino acids or modifications of similar masses (e.g., acetylation versus trimethylation, arginine versus dimethyl-lysine). Nevertheless, besides the used instrument, the search software and the search settings also have a very big impact on the database-dependent results of MS/MS spectra interpretation.

De Novo Sequencing

The database-dependent peptide identification strategy does not allow the identification of unexpected PTMs or unusual peptides that are not represented in protein databases. In some cases it may therefore be necessary to extract an amino acid sequence from an MS/MS spectrum in a fully database-independent manner. This is referred to as *de novo* peptide sequencing (Figure 3.2). Manual *de novo* sequencing is a tedious task that requires expertise and patience, and therefore several computer programs were developed that allow an automated assessment of a peptide's amino acids sequence. All *de novo* sequencing tools exclusively use the information in the MS/MS spectrum to derive an amino acid sequence. Several tools are now available and provide reliable *de novo* sequencing results with high-quality spectra [22–29]. All tools, however, suffer from inherent properties of MS/MS spectra that are often characterized by inaccurate measurements, missing peaks (gaps), and noise. Up-to-date software tools provide a probability estimate whether the extracted amino acid sequence is correct and may additionally assign probabilities to sequence substrings. The best-performing tools use probabilistic approaches and include, for example, PepNovo [28], PEAKS [24], and the Novo hidden Markov model (HMM) [27]. PepNovo is freely available and uses a probabilistic network whose structure reflects the physicochemical characteristics of peptide fragmentation in CID. PEAKS is a commercial software that computes the best possible peptide sequence for MS/MS spectra and provides confidence scores for amino acids in the sequence. Novo HMM generates a model that calculates emission probabilities for the suggested amino acid sequence from the observed spectrum. The last few years have seen significant improvements in MS methods and particularly in software tools to improve MS/MS spectrum analysis. These developments have now paved the way for the acquisition of reliable, high-quality peptide sequences in a database-independent way, which opens the road for the proteome analysis of organisms with unsequenced genomes.

Identification of PTMs

PTMs of proteins have fundamental functions in many cell processes (e.g., cell division, cell growth, and differentiation), and thus their identification is of great interest. The use of MS to identify modifications and also to identify the site of modification has become a very attractive method and was previously reviewed in Larsen et al. [30]. It is highly sensitive, and it has the ability to identify the site of modification and also to detect novel modifications. Nevertheless, good sample preparation and instruments with high sensitivity, high mass accuracy and resolution are required. PTMs of proteins are chemical groups that are covalently bound to specific amino acid residues by catalytic enzymes and result in mass increments as compared to the unmodified protein. Examples for such modifications are: phosphorylation, methylation, acetylation,

ubiquitination, and so on. MS analysis of proteins and peptides can reveal information on the type of modification; however, to assign the site of modification, it requires sequencing of the respective peptide by MS/MS. Modifications can be distinguished into stable and labile modifications. While stable modifications such as acetylation of Lys remain intact during CID, labile ones result in elimination of the group during fragmentation; for example, phosphorylation of Thr results in the loss of phosphoric acid (−98 amu).

Diagnostic ions and also neutral loss of chemical groups are modification specific signals that are generated during fragmentation. They are low-molecular-mass ions and are characteristic for a certain modification. Thus they also allow the distinction of modifications with similar mass—for example, Lys-acetylation and Lys-trimethylation. Acetylation and trimethylation give a mass increase of m/z 42.011 and 42.047, respectively. Fragmentation of peptides containing acetyl-Lys results in a characteristic immonium ion of m/z 126; however, trimethyl-Lys results in a neutral loss of m/z −59. These markers can therefore distinguish the two modifications. Table 3.3 provides a summary of some common modifications, along with their mass increments, neutral losses, and diagnostic ions.

For the analysis of PTMs, computational search algorithms exist that consider the Δ mass and sometimes also neutral loss for identification. However, there is a rate of false positives (FPs) that are annotated, and thus modifications identified by these search algorithms have to be analyzed manually in more detail. The presence of more than one modification on a single peptide can complicate the interpretation immensely and also increases the number of FP identifications in a database search.

To successfully map PTMs, sample preparation is very important. Usually, PTMs have low concentrations and thus the complexity of the sample has to be reduced or even better PTM-specific proteins or peptides have to be enriched. In phosphoproteomics, for example, phosphopeptides can be enriched by IMAC (immobilized metal-ion affinity chromatography) prior MS/MS analysis to increase the sensitivity.

TABLE 3.3. Summary of Mass Values, Diagnostic Ions, and Neutral Loss Ions of Some PTMs

Modification	D Mass (Da)	Amino Acid	Neutral Loss (Da)	Diagnostic Ions (Da)	References
Phosphorylation	79.966	Tyrosine, threonine, serine	79.966 97.977		30,31 31 31
Acetylation	42.0105	Lysine		126.091, 143.118	32,33
Monomethylation	14.0156	Lysine arginine		98.097	34 35,36
Dimethylation	28.0312	Lysine arginine		71.061, 46.066	34 35,36
Trimethylation	42.0468	Lysine	59.0735		34
Ubiquitination	114.1026	Lysine			37
Oxidation	15.9949	Methionine			38

The possibility to use the latest MS technologies, along with the further improvement of these techniques, will improve the mapping of PTMs and thus support the understanding of a lot of biological processes.

3.9 CONCLUSIONS

MS is a powerful tool in life sciences and thus also in plant biology. Proteome analysis by MS is a robust and fast technology giving information about protein composition, PTMs, localization, and quantity. Today a variety of mass spectrometers exists and it is not possible to judge which design is best. All of them have advantages and disadvantages. The choice of instrument is very much dependent on the type of experiment. All types of machines are continuously further developed to increase their performance and to circumvent disadvantages. The high mass accuracy and resolution makes FTICR-MS a very promising technology. However, its maintenance is very expensive and it also requires a lot of space. Therefore the latest ion traps, such as linear ion trap and the orbitrap, might be a cheaper and less space-using alternative. However, to study PTMs, not only high mass accuracy is important, but also good resolution in the LMW range is necessary, to detect diagnostic marker ions. In this range, Q and TOF analyzers are still important mass analyzers. Quadrupole is also still very advantageous for neutral loss scans to analyze, for example, phosphorylation. MS is becoming more and more automated, and the high throughput of samples yields huge amounts of data. Although several search algorithms exist and the output of good results is high, manual data interpretation is still required, especially to identify PTMs.

3.10 FIVE-YEAR VIEWPOINT

MS instruments are continuously evolving. Although MS instruments are already very sensitive and fast and have a high mass accuracy, there is still the potential to further improve these technical aspects in the near future. The high mass accuracy of FTICR instruments makes the relatively new field of top-down proteomics very attractive. Compared to bottom-up proteomics, top-down proteomics has the advantage to access complete protein sequences, to assign PTMs, and to prevent the protein inference problem, which results from the loss of connectivity between peptide and protein. However, this field still has some limitations that require further improvement—for example, interpretation of complex spectra.

REFERENCES

1. Barber M and Green BN. 1987. *Rapid Commun Mass Spectrom* **1**:80–83.
2. Karas M and Hillenkamp F. 1988. *Anal Chem* **60**:2299–2301.
3. Fenn JB, et al. 1989. *Science* **246**:64–71.

4. Yost RA and Boyd RK. 1990. *Methods Enzymol* **193**:154–200.
5. Loboda AV, et al. 2000. *Rapid Commun Mass Spectrom* **14**:1047–1057.
6. Shevchenko A, et al. 2000. *Anal Chem* **72**:2132–2141.
7. March RE. 1997. *J Mass Spectrom* **32**:351–369.
8. Jonscher KR and Yates JR 3rd. 1997. *Anal Biochem* **244**:1–15.
9. Alan G and Marshall C. 1998. *Mass Spectrom Rev* **17**:1–35.
10. Hu Q, et al. 2005. *J Mass Spectrom* **40**:430–443.
11. Hardman M and Makarov AA. 2003. *Anal. Chem.* **75**:1699–1705.
12. Chen S and Harmon AC. 2006. *Proteomics* **6**:5504–5516.
13. Rose JK, et al. 2004. *Plant J* **39**:715–733.
14. Griffin PR, et al. 1995. *Rapid Commun Mass Spectrom* **9**:1546–1551.
15. Pappin DJ, Hojrup P, and Bleasby AJ. 1993. *Curr Biol* **3**:327–332.
16. Perkins DN, et al. 1999. *Electrophoresis* **20**:3551–3567.
17. Eng JK, McCormack AL, and Yates JR 3rd. 1994. *J Am Soc Mass Spectrom* **5**:976–989.
18. Colinge J, et al. 2004. *Proteomics* **4**:1977–1984.
19. Colinge J, et al. 2005. *Anal Chem* **77**:596–606.
20. Roepstorff P and Fohlman J. 1984. *Biomed Mass Spectrom* **11**:601.
21. Biemann K. 1990. *Methods Enzymol* **193**:886–887.
22. Chen T, et al. 2001. *J Comput Biol* **8**:325–337.
23. Johnson RS and Taylor JA. 2002. *Mol Biotechnol* **22**:301–315.
24. Ma B, et al. 2003. *Rapid Commun Mass Spectrom* **17**:2337–2342.
25. Zhang Z. 2004. *Anal Chem* **76**:6374–6383.
26. Searle BC, et al. 2004. *Anal Chem* **76**:2220–2230.
27. Fischer B, et al. 2005. *Anal Chem* **77**:7265–7273.
28. Frank A and Pevzner P. 2005. *Anal Chem* **77**:964–973.
29. Grossmann J, et al. 2005. *J Proteome Res* **4**:1768–1774.
30. Larsen MR, et al. 2006. *Biotechniques* **40**:790–798.
31. Steen H, et al. 2001. *Anal Chem* **73**:1440–1448.
32. Neubauer G and Mann M. 1999. *Anal Chem* **71**:235–242.
33. Zhang K, et al. 2002. *Anal Biochem* **306**:259–269.
34. Kim JY, et al. 2002. *Anal Chem* **74**:5443–5449.
35. Zhang K, et al. 2004. *Proteomics* **4**:1–10.
36. Gehrig PM, et al. 2004. *J Am Soc Mass Spectrom* **15**:142–149.
37. Rappsilber J, et al. 2003. *Anal Chem* **75**:3107–3114.
38. Peng J, et al. 2003. *Nat Biotechnol* **21**:921–926.

CHEMICAL PROTEOMICS

Miriam C. Hagenstein, Olaf Kruse, and Norbert Sewald

4.1 INTRODUCTION

Despite the tremendous advancements of proteomics technologies that have been made during the past years, several challenges remain to be tackled. A eukaryotic cell may produce more than 10,000 different proteins at a time with a high dynamic concentration range. Consequently, both the separation and analysis of such a huge number of proteins is difficult, especially when it comes to more complex systems. Moreover, because many proteins and protein classes of high interest are not highly abundant in a biological system, but are instead present in very low numbers, either depletion strategies of highly abundant proteins are necessary or enrichment/detection strategies for the target proteins of interest need to be developed.

4.2 STRATEGIES FOR ACTIVITY-BASED PROTEIN PROFILING (ABPP)

The family-specific detection or enrichment of protein classes based on a common binding mechanism has been proven feasible by using small-molecule ligands that

address protein families in a semispecific manner [1, 2]. Such enrichment strategies have been termed *chemical proteomics* or *functional proteomics* because they are based on chemically synthesized molecular tools and address specific functions of the target proteins. Because binding of (for example) enzyme inhibitors to their target protein relies on the activation state of the enzyme, such approaches have also been named ABPP or *activity-based proteomics*.

Two different groups of target proteins that are being addressed by distinct types of small-molecule ligands with consequences for the design of the molecular tools necessary for the application in functional proteomics can be distinguished: (i) proteins (enzymes) where irreversibly binding ligands or suicide substrates are known and (ii) proteins (not necessarily enzymes) where only reversibly binding ligands (including co-factors and their mimetics) are known. The target protein family that is being addressed by the protein ligand has to be labeled with a reporter group, as schematically shown in Figure 4.1. For the first class of proteins where irreversibly binding ligands are known, the two properties *molecular recognition* and *chemical reactivity* are both present in the irreversible inhibitor, while for the second class of proteins, where only reversibly binding ligands as recognition units are known, an additional molecular functionality in the form of a reactive group has to be introduced into the activity-based probes. In addition, the protein ligands have to be linked to a reporter tag (biotin, fluorescence label, or radioactive labels). Alternatively, a subproteome generation aiming at an enrichment of the protein class which is being addressed by reversibly binding inhibitors can be achieved by inhibitor affinity chromatography (IAC). For that purpose the inhibitor or generally the protein ligand is chemically modified in such a way that it can be immobilized to a solid surface [affinity chromatography matrix, magnetic beads, or surface plasmon resonance (SPR) sensor chips]. Several strategies have been developed on the basis of the above-mentioned requirements.

Directed Affinity-Based Probes (ABP) Using Irreversible Inhibitors

Conjugates of irreversibly binding ligands (inhibitors) with reporter groups are appropriate molecular tools for functional proteomics in order to retrieve many or all

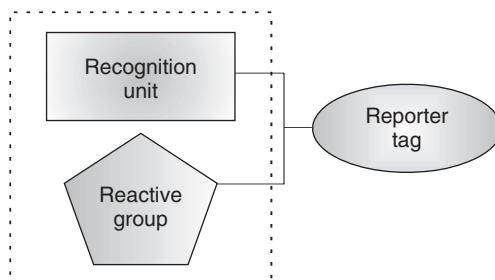


FIGURE 4.1. Activity-based probe. Irreversible inhibitors are characterized by combining “recognition unit” and “reactive group,” whereas molecular tools based on reversible inhibitors require an additional reaction site for covalent cross-linking.

representatives of a class of proteins in a family-specific manner. Such irreversible inhibitors comprise both the recognition unit and the reactive group in one molecular entity and can be divided into two subclasses encompassing either reactive mechanism-based probes or suicide substrate probes [3]. These two groups differ in such a way that the latter one is being transformed into a highly reactive species in the course of the enzyme reaction which is then able to form a covalent bond to the active site of the enzyme. In the other case the probe intrinsically displays high reactivity toward nucleophiles. In both cases the molecular probe is directed toward enzymatically active proteins because only such species are being addressed by the inhibitor or suicide substrate. Consequently, such molecular tools can be used for ABPP. Representative pairs of protein families and irreversibly binding ligands are compiled in Table 4.1.

Fluorophosphonates address serine proteases [4] as they react with the catalytic Ser residue in the active site. Such Ser residues are present not only in serine proteases but also, in general, in serine hydrolases, which form one of the largest enzyme families present in eukaryotes. Consequently, such affinity-based probes are of high relevance to functional proteomics both in plants and animals. It was estimated that serine hydrolases represent approximately 1% of the predicted protein products encoded by the human genome. In the context of plant proteomics, such probes are suited to address lipid hydrolases, esterases, and amidases. In a similar fashion, cysteine proteases may be targeted with epoxy succinyl moieties [5, 6], vinyl sulfones [7–9], acyloxymethyl ketones [10, 11], acrylates [12], and fluoromethyl ketones [13] (Table 4.1). All compounds have in common that they are potent electrophiles that can be attacked by thiol groups under physiological conditions.

Additional recognition units present in the affinity-based probe such as, for example, peptides or peptide-like linkers confer improved selectivity toward certain classes of proteases. Acyloxymethylketones that comprise an Asp residue (Table 4.1) close to the probe head turned out to be selective for caspases [10], while a library of peptide vinyl sulfones (PVS, Table 4.1) with varying amino acids in the linker has been employed to profile substrate and inhibitor selectivity of the catalytic subunits of the proteasome [7]. Solid-phase synthesis of PVS permits the rapid optimization of linker length and properties as it was shown by Overkleeft et al. [14].

Acyloxymethyl ketones [10, 11] are characterized by low reactivity toward weak nucleophiles but readily react with thiol residues in the active site of enzymes. Acrylates and vinylogous amino acids (Table 4.1) proved to be appropriate for targeting cysteine proteases [12]. In particular, approaches directed toward caspase subproteome generation suffered from shortcomings: Several currently available activity-based probes have the limitation of a high level of background labeling when applied to crude proteomes. Moreover, not all caspases can be addressed in a similar manner because they likewise react readily with thiol nucleophiles. In particular, fluoromethyl ketones served for the synthesis of activity-based probes addressing caspases [13].

The ABP directed toward protein tyrosine phosphatases (PTPs) either rely on the reaction of an electrophilic probe with an active site Cys residue (Table 4.1) [15, 16] or on the incorporation of a phosphatase suicide inhibitor (Table 4.2). The tyrosine phosphatase hydrolyzes the phenyl phosphate releasing a phenoxide that contains an

TABLE 4.1. Directed Affinity-Based Probes Employing Irreversible Inhibitors

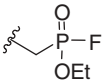
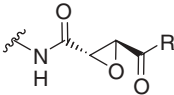
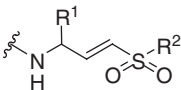
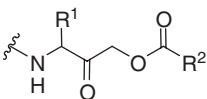
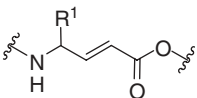
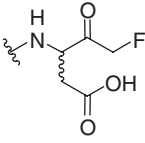
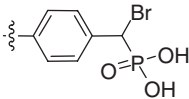
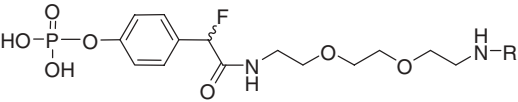
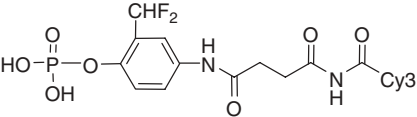
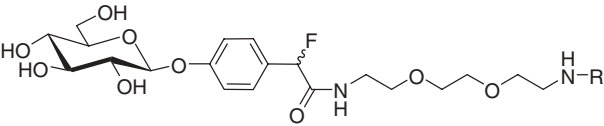
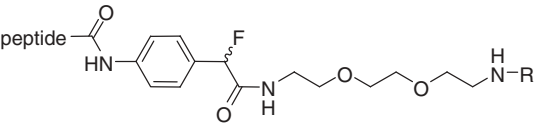
Target Enzymes	Reactive Group		References
Serine hydrolases	Fluorophosphonate		4
Cysteine proteases	Epoxy succinyl moiety		5, 6
Cysteine proteases	Vinyl sulfone		7–9
Cysteine proteases	Acyloxymethylketone		10, 11
Cysteine proteases	Acrylate, vinylogous amino acid		12
Cysteine proteases	Fluoromethylketone		13
Tyrosine phosphatases	α -Bromobenzylphosphonate		15, 16

TABLE 4.2. Molecular Tools Employing Suicide Substrates of Phosphatases, Glycosidases, and Proteases

Target Enzymes	Suicide Substrate	References
Tyrosine phosphatases		17, 18
Tyrosine phosphatases		17
Glycosidases		19
Proteases		20

additional leaving group (e.g., fluoride) in an appropriate position. Fluoride elimination results in the formation of a potent quinone methide electrophile [17, 18]. However, the reaction of the quinone methide with nucleophiles is rather slow, which permits diffusion of the reactive species out of the active site. Consequently, these ABPs must be used in a rather high concentration. The application of these molecular tools in complex proteomes remains to be proven.

The principle of hydrolysis of a strategic bond followed by elimination to create a reactive species has been employed also for glycosidases [19] and proteases [20]. The selectivity of the cleavage reaction and consequently the selection of the target protein class are defined by the glycosyl residue or the peptide (Table 4.2). A naturally occurring general inhibitor of cysteine proteases (E-64, Figure 4.2) contains an epoxy succinyl moiety and serves as a blueprint of such inhibitors while modification of the conjugates provided affinity labels for the papain subclass [21]. Combination of the structural and binding characteristics of E-64 and CA074 resulted in the highly potent and selective chimeric inhibitor NS-134 targeting the cysteine protease subclass of the cathepsins [5].

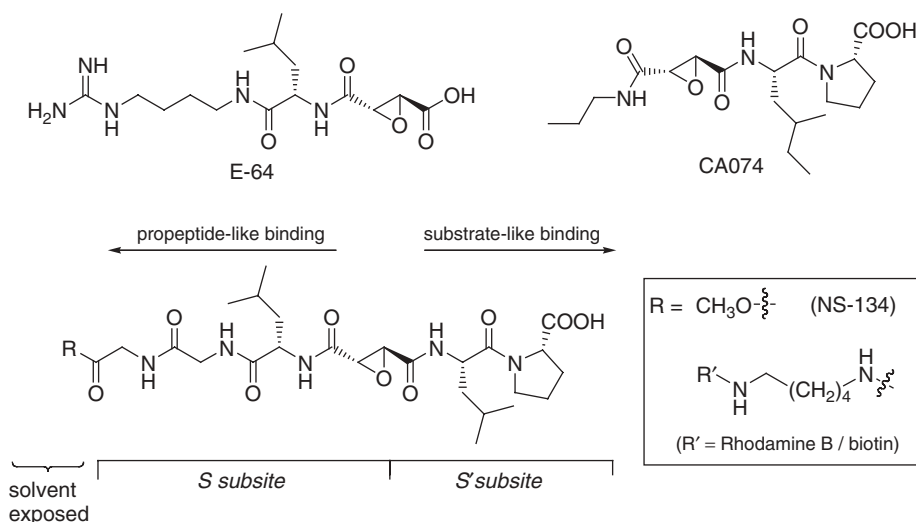


FIGURE 4.2. Epoxysuccinyl-type ABPs derived from naturally occurring protease inhibitors E-64 and CA074.

Reversibly Binding Ligands as Recognition Units

Enrichment by IAC. Enrichment of the target proteins by IAC is a viable alternative for the majority of proteins, because inhibitors or irreversibly binding ligands are not known. Generation of a proteome subset by IAC succeeds by immobilization of small organic compounds (e.g., engineered inhibitors) on an affinity chromatography column. This has been successfully proven in the creation of metalloprotease subproteomes [22, 23] or kinase subproteomes [24–26]. For the optimization process of the elution conditions in the IAC process, SPR represents an excellent method [27].

Labeling Strategies Involving Reversible Inhibitors. Whenever labeling of a proteome subset on a common mechanistic basis using reversibly binding ligands is desired, an additional chemical step is necessary to obtain a covalent bond between the probe and the target protein. This is the major precondition for subproteome analysis according to the work flow comprising 2-DGE and PMF analysis. Consequently, novel synthetic molecular tools are required that are able to address the target protein family and to undergo a cross-linking reaction upon external triggering. Such a cross-linking reaction can be brought about photochemically: Photoaffinity labeling has been utilized in a manifold manner for localization of ligand-binding sites in biochemistry [28]. There are different photoreactive groups available that can be incorporated into ABPs.

The incubation of a proteome with an ABP comprising a recognition unit, a reporter group, and a photoreactive probe, followed by photochemical cross-linking, has been proven suitable to retrieve target proteins that bind to the recognition unit of the molecular tool. This concept is suited for many different classes of proteins and may facilitate the discovery of new members of a protein family. In such an approach

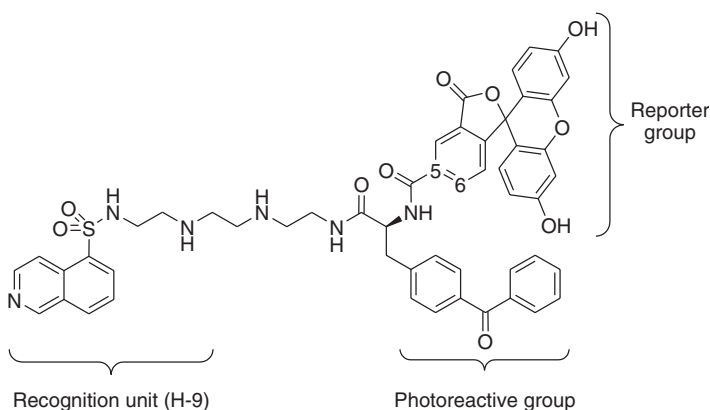


FIGURE 4.3. Derivative of the isoquinoline sulfonamide-type kinase inhibitor H-9, equipped with a photoreactive group and a reporter group (fluorophor) [29].

a proteome subset is being labeled with high sensitivity and can then be subjected to further analysis. Although FP results cannot be excluded, this method is appropriate to significantly reduce the number of proteins in a complex proteome. The application of such tools for quantification would require the photoreaction to proceed quantitatively, which is not at all realistic. Nevertheless, such probes can be employed efficiently for qualitative analysis.

Labeling of kinases (*vide infra*, *Case Study*) with a fluorophor has been shown to be possible [29] with a conjugate that is composed of a broad-spectrum kinase inhibitor (e.g., the isoquinoline sulfonamide H-9), *p*-benzoylphenylalanine (Bpa) as the photoreactive group, and carboxyfluorescein as the fluorescent label. The basic concept, pioneered by Hagenstein et al. [29], has subsequently been utilized and validated for (for example) metalloproteases [27, 30–32] by conjugating the Zn-chelating hydroxamate inhibitors to a photoreactive group and different reporter groups (Table 4.3).

Besides benzophenone-type photoreactive groups [28, 30] a diazirine [32] was also used (Table 4.3). Chattopadhyaya et al. [33] extended the successful application of such an approach toward aspartyl proteases (Table 4.3), while a similar probe [34] was employed for tagging of carbohydrate binding proteins (lectins). In that case a sugar moiety was employed as a ligand; and instead of attaching directly a reporter group, an azide moiety was incorporated. This chemical functionality subsequently allows in a completely orthogonal way to attach reporter groups after subproteome generation in the frame of click chemistry.

4.3 CASE STUDY: DEVELOPMENT OF MOLECULAR TOOLS TARGETING PLANT KINASES

For protein profiling by using reversibly binding ligands, such a moiety has to be equipped with a reporter group and an additional reactive group (Figure 4.1). By

analogy to photoaffinity labeling [28], the conjugate is linked covalently to the protein in the course of a photoreaction. This avoids dissociation under the conditions of 2-DGE. The new concept has been proven feasible for the labeling of kinases [29].

Kinases play a major role in many regulatory mechanisms of all living cells. Phosphorylation/dephosphorylation processes trigger important metabolic pathways and participate in internal and external adaptation mechanisms. Systematic analysis of completely sequenced genomes provided crucial advances in kinase research. Although progress has been made in understanding detailed functions of plant kinases, it is still very difficult to assign the vast amount of recently identified kinase genes by systematic genomics (more than 196 in *A. thaliana*) to a particular cell regulation process. Ser/Thr kinases, of particular interest in plants, make up more than 12 groups because of sequence relationships. Thylakoid membranes contain Ser/Thr kinases functionally involved in adaptation mechanisms as biological responses to changes in environmental conditions (e.g., light) [35–37]. Mapping of *in vivo* protein phosphorylation sites of surface-exposed phosphorylated peptides in photosynthetic membranes of the green alga *C. reinhardtii* has recently been studied in detail by IMAC in combination with nESI–Q–TOF–MS/MS sequence analysis [38]. A total of 19 *in vivo* phosphorylation sites were mapped in the proteins corresponding to 15 genes in *C. reinhardtii* [38], with many of them belonging to light harvesting complex (LHC) proteins and photosystem II (PS II) proteins. Protein phosphorylation of these proteins is highly sensitive to inhibitors of cAMP-dependent kinases [39, 40]. Some of these kinases have been recently identified as thylakoid-associated enzymes in *A. thaliana* and in *C. reinhardtii* using forward and reverse genetic approaches [41–44]. Two of them, the LHC kinases *STN7* in *Arabidopsis* and *STT7* in *Chlamydomonas* [41, 42], are sensitive to the isoquinoline sulfonamide-type inhibitor H-9, which interacts with their ATP binding site [43].

A chemical probe for kinases was designed on the basis of H-9, with a photoreactive group and a fluorophor being connected to the reversibly binding recognition unit H-9 across a polar linker. This conjugate aims at tagging Ser/Thr kinases by photochemical cross-linking followed by 2-DGE. The attachment site of the linker was selected on the basis of the crystal structure of a complex between a H-9 derivative and a kinase [45]. 4-Bpa was employed as the photoreactive group, because it is stable to common protic solvents and brings about efficient single-site modification [28] upon photochemical triplet insertion into CH bonds after irradiation at 350 nm.

A series of commercially available kinases was subjected to an *in vitro* assay in order to examine the specificity of photoaffinity labeling with the photoreactive and fluorescent H-9 conjugate. Three kinases (Figure 4.4, lanes 1, 2, and 4) were fluorescently tagged to a high extent, whereas the labeling efficiency of a fourth one (lane 3) was quite poor. The specificity of the photoaffinity labeling process is emphasized by the fact that, for example, only the catalytic subunit of cAMP-dependent kinase was tagged, while the regulatory units remained unlabeled. The qualitative data were supported in real-time binding studies using SPR [29]. Potential radiationless fluorescence resonance energy transfer (FRET) within the conjugate—for example, between the benzophenone triplet state and the neighboring fluorescein—could be ruled out according to excitation-emission spectroscopy (EES) [29]. Energy transfer between the

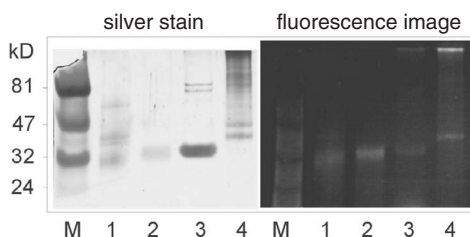


FIGURE 4.4. Photoaffinity labeling of kinases with the H-9 derivative occurs specifically: Lane 1: hexokinase. Lane 2: Creatin kinase. Lane 3: 3PGA phosphokinase. Lane 4: cAMP-dependent protein kinase C. Lane M: Protein marker. (Figure reproduced from reference 29.)

isoquinoline of H-9 and fluorescein was also found to be negligible. Consequently, quenching and chemical reactions are the dominant loss channels after photoactivation at 350 nm.

The native folded form of the protein was shown to be a prerequisite for recognition by the protein ligand and efficient photoaffinity labeling. With denatured kinases, no photo-cross-linking occurred: Denaturation of the proteins with SDS prior to incubation and photoreaction did not result in fluorescent bands after gel electrophoresis (Figure 4.5, lanes 2, and 4), which proves that the intact kinase is required for labeling.

The genome of *C. reinhardtii* had been sequenced but not yet been annotated completely at the time of the experiments, which precludes efficient mass spectrometric fingerprint analysis after tryptic digestion. Human creatine kinase was spiked into a thylakoid protein preparation from *C. reinhardtii* prior to the labeling step to allow for further validation of the chemical proteomics method. The spiked creatine kinase could be retrieved (Figure 4.6, lane 4) and identified unambiguously by MALDI-TOF-MS

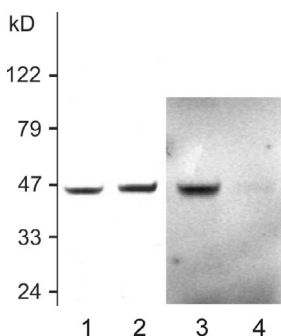


FIGURE 4.5. Selective kinase tagging by photoaffinity labeling with the H-9 derivative. Comparison of photoaffinity labeling efficiency of native creatine kinase with labeling of denatured creatine kinase. Lane 1: Silver stain, native creatine kinase. Lane 2: Silver stain, denatured creatine kinase. Lane 3: Fluorescence image, native creatine kinase. Lane 4: Fluorescence image, denatured creatine kinase. (Figure reproduced from reference 29.)

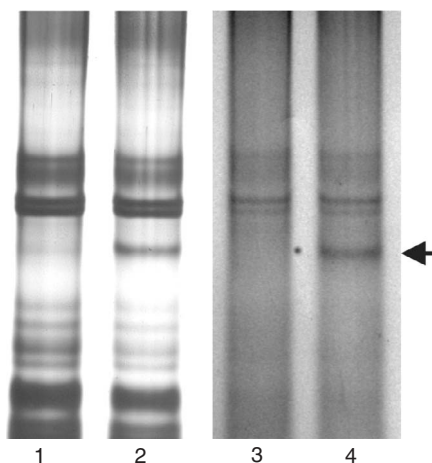


FIGURE 4.6. Retrieval of spiked creatine kinase from thylakoid membrane preparations after photoaffinity labeling. Lane 1: Silver stain. Lane 2: Silver stain, with additional 150 ng creatine kinase. Lane 3: Fluorescence image. Lane 4: Fluorescence image, with additional 150 ng creatine kinase (see arrow). (Figure reproduced from reference 29.)

fingerprint analysis. The H-9 conjugate also tagged several thylakoid membrane proteins predominantly in the mass range between 40 kDa and 80 kDa (Figure 4.6, lane 3). It is noteworthy that the LHC proteins (Figure 4.6, proteins with masses < 33 kDa) are not labeled, which underscores the selectivity of the labeling process.

4.4 CONCLUSIONS

Several sophisticated methods have been invented in the past to enrich or globally monitor protein families as well as changes in protein expression level under different physiological conditions. Activity-based chemical proteomics provides tailor-made molecular tools and, hence, is a powerful method to enrich, tag, or characterize protein families, especially those of low abundance and with distinct functions in different cellular compartments. ABPs comprising irreversible inhibitors have been proven to be suitable molecular tools for the tagging of enzyme targets but also for profiling of enzyme activity. Probes relying on reversibly binding ligands may be employed either for inhibitor affinity purification or—in combination with an additional reactive site (e.g., a photoreactive group)—for tagging the target proteins to suitable reporter groups. Proteins for which irreversible inhibitors are unknown can be addressed with this latest generation of ABPs. Kinases are a typical example for such large groups of protein families with functions in signaling and sensing of physiological processes. In plant proteomics, it is of particular interest to identify subfamilies of kinases in chloroplasts, mitochondria, and the cytosol as a first step to elucidate their role in environmental adaptation processes and in inter-organelle communication. The first successful steps toward the semiquantitative tagging and identification of Ser/Thr

protein kinases within chloroplasts of *Chlamydomonas* underscore the impact of this method for elucidating their cellular function.

4.5 FIVE-YEAR VIEWPOINT

A basic set of novel molecular tools has been developed during the past few years. They are suited for the enrichment or detection of mechanistically related proteins. The scope of targets has been extended from enzymes to the full proteome by using protein ligands instead of enzyme inhibitors. Without any doubt, the scope of the target proteins to be analyzed will be widened during the next five years. New probes with more ligands need to be developed. Chemical proteomics will be developed into a new platform technology, where the basic principles shown in this chapter will be utilized for high-throughput analysis in highly sophisticated devices. Further developments of chemical proteomics can be seen in the field of interaction analysis (interactomics). Moreover, the application in a living system will finally allow study of the interaction and trafficking of biomolecules in a cell.

REFERENCES

1. Hagenstein MC and Sewald N. 2006. *J Biotechnol* **124**:56–73.
2. Evans MJ and Cravatt BF. 2006. *Chem Rev* **106**:3279–3301.
3. Jeffrey DA and Bogoy M. 2003. *Curr Opin Biotech* **14**:87–95.
4. Liu Y, Patricelli MP, and Cravatt BF. 1999. *Proc Natl Acad Sci USA* **96**:14694–14699.
5. Schaschke N, et al. 2000. *FEBS Lett* **482**:91–96.
6. Greenbaum D, et al. 2002. *Mol Cell Proteomics* **1**:60–68.
7. Nazif T and Bogoy, M. 2001. *Proc Natl Acad Sci USA* **98**:2967–2972.
8. Wang G, et al. 2003. *Org Lett* **5**:737–740.
9. Ovaa H, et al. 2003. *Angew Chem Int Ed Engl* **42**:3626–3629.
10. Thornberry NA, et al. 1994. *Biochemistry* **33**:3934–3940.
11. Kato D, et al. 2005. *Nature Chem Biol* **1**:33–38.
12. Winssinger N, et al. 2002. *Proc Natl Acad Sci USA* **99**:11139–11144.
13. Liao M-L, Panicker RC, and Yao SQ. 2003. *Tetrahedron Lett* **44**:1043–1046.
14. Overkleeft HS, et al. 2000. *Tetrahedron Lett* **41**:6005–6009.
15. Kumar S, et al. 2004. *Proc Natl Acad Sci USA* **101**:7943–7948.
16. Kumar S, et al. 2006. *J Proteome Res* **5**:1898–1905.
17. Zhu Q, et al. 2003. *Tetrahedron Lett* **44**:2669–2672.
18. Lo L-C, et al. 2002. *J Proteome Res* **1**:35–40.
19. Tsai C-S, Li Y-K, and Lo L-C. 2002. *Org Lett* **4**:3607–3610.
20. Zhu Q, et al. 2004. *Chem Commun* **13**:1512–1513.
21. Greenbaum D, et al. 2000. *Chem Biol* **7**:569–581.
22. Freije JR and Bischoff R. 2003. *J Chromatogr A* **1009**:155–169.
23. Collet MI, et al. 2007. *J Biotechnol* **129**:316–328.

24. Lolli G, et al. 2003. *Proteomics* **3**:1287–1298.
25. Wissing J, et al. 2004. *Mol Cell Proteomics* **3**:1181–1193.
26. Brehmer D, et al. 2004. *Mol Cell Proteomics* **3**:490–500.
27. Jenssen K, Sewald K, and Sewald N. 2004. *Bioconjug Chem* **15**:594–600.
28. Dorman G and Prestwich GD. 2000. *Trends Biotechnol* **18**:64–77.
29. Hagenstein MC, et al. 2003. *Angew Chem Int Ed Engl* **42**:5635–5638.
30. Jenssen K. 2004. Peptide und peptidomimetische Verbindungen als Werkzeuge in der Proteomanalyse, Ph.D. Thesis, Bielefeld University.
31. Saghatelian A, et al. 2004. *Proc Natl Acad Sci USA* **101**:10000–10005.
32. Chan EWS, et al. 2004. *J Am Chem Soc* **126**:14435–14446.
33. Chattopadhyaya S, Chan EWS, and Yao SQ. 2005. *Tetrahedron Lett* **46**:4053–4056.
34. Ballell L, et al. 2005. *Chem Bio Chem* **6**:291–295.
35. Gal A, Zer H, and Ohad I. 1997. *Physiol Plant* **100**:869–885.
36. Allen JF. 1992. *Biochim Biophys Acta* **1098**:275–335.
37. Kruse O. 2001. *Naturwissenschaften* **88**:284–292.
38. Turkina MV, et al. 2006. *Mol Cell Proteomics* **5**:1412–1425.
39. Ishak Z, White IR, and Millner PA. 1992. *Biochem Soc Trans* **20**:8S.
40. Farchaus J, Dilley RA, and Cramer WA. 1985. *Biochim Biophys Acta* **809**:17–26.
41. Bellafiore S, et al. 2005. *Nature* **433**:892–895.
42. Depege N, Bellafiore S, and Rochaix JD. 2003. *Science* **299**:1572–1575.
43. Bonardi V, et al. 2005. *Nature* **437**:1179–1182.
44. Snyders S and Kohorn BD. 2001. *J Biol Chem* **276**:32169–32176.
45. Eng R, et al. 1996. *J Biol Chem* **271**:26157–26164.

THE *ARABIDOPSIS* LOCALIZOME: SUBCELLULAR PROTEIN LOCALIZATION AND INTERACTIONS IN *ARABIDOPSIS*

Georgios Kitsios, Nicolas Tsesmetzis, Max Bush,
and John H. Doonan

5.1 PROTEIN COMPARTMENTALIZATION IN PLANT CELLS

Like all eukaryotic cells, plant cells are organized into structurally and functionally distinct compartments whose organization and composition can be highly dynamic. Early cytological studies using simple visual inspection under the light microscope indicated that the interior of cells was not uniform and revealed various subcellular features such as the nucleus, vacuole, and chloroplasts. Moreover, these subcellular structures were mobile, and they changed shape and divided. The application of chemical dyes that differentially stained different parts of the cell supported the notion that these compartments differed in chemical composition and therefore were probably functionally distinct. The development of the electron microscope (EM) and its application to biological specimens revealed the immense internal complexity of cells. Thin sections of resin-embedded cells could be stained with heavy-metal-containing compounds to provide clear views of the cell's substructure. Membrane-bound compartments were clearly observed and included the nucleus, chloroplasts, and mitochondria as well as a diverse collection of less well-defined structures. The presence of a bounding

membrane indicated that communication between the inside and the outside of these organelles might be restricted. Highly structured pores were evident in the nuclear envelope and suggested a route for nuclear–cytoplasmic transport, but the mechanisms for other organelles were less clear. EM also revealed other regions, such as the nucleolus, that clearly differed in structure from neighboring parts of the cell but was not bounded by a membrane. The organization of such regions is presumably maintained by other mechanisms and may depend on the underlying function of that region. For instance, the structure of the nucleolus is intimately linked to the production of ribosomes [1, 2].

Changes in compartment organization and shape can be mediated by a number of mechanisms, some of which are specific for the particular compartment. A fibrous system of filamentous proteins extends throughout the cytoplasm and provides a scaffold that organizes and integrates overall cellular organization and transport within and between cells. These fiber-like structures, later to be collectively known as the cytoskeleton, appeared in early EM studies. Microtubules and actin microfilaments form two distinct but interacting fibrous systems that are found in almost every eukaryotic cell. Microtubules are composed of tubulin dimers, and they play direct roles in controlling the direction of cell expansion. Actin microfilaments are required for intracellular transport. These two systems are able to grow, shrink, break, cross-link, and form bundles or interconnect into networks [3], and much of their function derives from this dynamic ability to reorganize. Biochemical and genetic characterization of cytoskeletal proteins reveals tremendous specialization into groups with primarily structural functions, organization, or motor functions. Filamentous proteins and their associated motors, therefore, provide a dynamic scaffold to drive both protein and compartment mobility.

This complex substructure of the cell raised the question as to how proteins reach their location. A major breakthrough was made by Blobel when his group showed that many proteins contained “address tags,” strings of amino acids that directed a protein to a particular destination [4]. There is often a close relationship between subcellular location and function. In the case of structurally related proteins, differential subcellular localization can indicate specialized functions [5], and many proteins are regulated by sequestration in compartments away from their targets and substrates; thus a transcription factor (TF) retained in the cytoplasm is unlikely to be active. Therefore, subcellular location is a key characteristic of a protein and, along with primary sequence and bioinformatic analysis, can provide insight into its functional characteristics [6]. Large-scale genome sequencing over the past few years has revealed something of the diversity of proteins potentially present in cells. Entire genome sequencing is now complete for *Arabidopsis* (*Arabidopsis* Genome Initiative, 2000), rice (International Rice Genome Sequencing Project: IRGSP, 2005), and populus (http://genome.jgi-psf.org/Poptr1_1/) while extensive expressed sequence tags (ESTs) are available for many other plant species (NCBI Entrez Genome: <http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html>). This immense resource provides the basis to systematically define protein localization using two distinct but mutually supportive approaches, one predictive and the other experimentally driven.

Predicting Protein Localization

Since many proteins may contain within their amino acid sequences “hidden” information that could be used for predicting their location within the cell, computational programs have been developed with the aim of revealing this and thereby predicting the location of the protein (see Table 5.1 for some examples). Simple linear sequences implicated, for example, in nuclear localization can be predicted with reasonable accuracy. However, localization predication can be complicated and confounded by superimposition of regulatory features: many proteins are retained in a compartment where they are inactive (due to lack of co-factors or substrates, for example) until specific requirements are met. Other targeting sequences are not simple and involve PPI, selective destruction due to compartment specific proteases, and other, less-well-defined mechanisms.

The experimental determination of localization now extends to many hundreds of proteins in plants and animals, and these data can be used to design predictive algorithms based on primary amino acid sequence or other structural features of potential proteins. Along with “molecular function” and “biological process,” “protein localization” or “cellular component” is used to classify gene products into groups (Gene Ontology (GO) Consortium, 2001). The cellular component encompasses the subcellular structures, locations, and macromolecular complexes in which a protein may be found. Knowledge of the cellular component (or subcellular location) of a gene product can be a strong indicator of the molecular function and/or the biological process. However, few systematic studies have been performed that concentrate on the cell component, and, as a result, the databases are populated with almost exclusively predicted data. This is especially so in plants where studies are more labor-intensive, more time-consuming, and significantly less efficient than those using mammalian cells or yeast [6–14].

It is therefore important to understand the uses and limitations of the computational tools that have been developed over the last 10 years. In contrast to experimental studies, such predictive tools provide an accurate subcellular localization prediction for a relatively small fraction of cases whereas dynamics and behavioral diversity are often not captured at all. Improvement in accuracy remains a goal for the bioinformatics community. However, five conceptually different approaches/concepts have made significant contributions [15]. These are based on: localization motif knowledgebases [16, 17], support vector machines [18–21], artificial neural network (ANN) [22–24], HMM [25], and a covariant discriminant algorithm (CDA) [26–28].

Localization Motif Knowledgebase (LMK)

LMKs search for local sequence motifs that are responsible for directing the protein to the right compartment. In contrast to methods that are based on amino acid composition, LMKs identify linear peptide sequences that are predicted to be responsible for localization. However, LMK algorithms are highly dependent on the prior identification of localization motifs (usually by experimental means) and, overall, they underperform compared to other prediction methods.

TABLE 5.1. Plant Protein Localization Predictors

Name	URL	Compartments					Method	Reference						
BaCellLo	http://gpccr.biocomp.unibo.it/bacello/	ch	c	e	m	n	SVM	33						
ChloroP	http://www.cbs.dtu.dk/services/ChloroP/	ch					ANN	23						
CoupleLoc	http://www.bioinfo.tsinghua.edu.cn/CoupleLoc/	ch	c	e	g	m	n	p	SVM	21				
iPSORT	http://hc.ims.u-tokyo.ac.jp/iPSORT/	ch		e	m				Decision list	154				
LOCtree	http://cubic.bioc.columbia.edu/services/loctree/	ch	c	e	m	n			SVM	20				
PA-SUB	http://www.cs.ualberta.ca/~bioinfo/PA/Sub/	ch	c	er	e	g	m	n	p	HMM	25			
PLOC	http://www.genome.jp/SIT/ploddir/	ch	c	cy	er	e	g	m	n	p	SVM	19		
Predotar	http://urgi.versailles.inra.fr/predotar/	ch		er	m					ANN	22			
Protein Prowler	http://pprowler.imb.uq.edu.au/	ch		e	m					ANN, SVM	155			
pSLIP	http://pslip.bii.a-star.edu.sg/	ch	c	e	m	n		pm		SVM	18			
TargetP	http://www.cbs.dtu.dk/services/TargetP/	ch		e	m					ANN	24			
WoLF PSORT	http://wolfpsort.org/	ch	c	cy	er	e	g	m	n	p	pm	vm	LMK	17
CDA	Not available	ch	c	cy	er	e	g	m	n	p	pm	v	CDA	26

ch, chloroplasts; c, cytosol; cy, cytoskeleton; er, endoplasmic reticulum; e, extracellular; g, golgi; m, mitochondria; n, nucleus; p, peroxisomes; pm, plasma membrane; v, vacuole; vm, vacuolar membrane; SVM, support vector machines; ANN, artificial neural network; HMM, hidden markov model; LMK, localization motif knowledgebase; CDA, covariant discriminant algorithm.

Support Vector Machine (SVM)

SVMs are a set of methods used for data classification. Using SVMs, input data can be projected in the form of vectors onto a high-dimensional space where a maximal separating hyperplane is constructed. Then, on each side of the hyperplane, a new hyperplane is constructed which separates the data. The best separating hyperplane is the hyperplane that maximizes the distance between the two newly created hyperplanes. Algorithms based on SVMs are very popular due to their robustness (especially when datasets are noisy), their high success rates, and the fact that they are less prone to overfitting.

Artificial Neural Network (ANN)

ANNs are nonlinear statistical data modeling tools that can be used to model complex relationships between inputs and outputs or to find patterns in data. Predictors based on ANNs are also very popular due to their high prediction accuracy, which is partly based on their ability to generalize and deal with noisy data if properly trained [29]. Unlike SVMs, ANNs are quite prone to “overfitting.” In other words, ANNs can be “tuned” so that they can make a flawless prediction of their training data, but because of that they lose their ability to generalize and therefore they may fail to interpret novel patterns in data unrelated to their training set. ANNs are regarded as “black boxes” since they cannot provide further interpretation of the results due to the inherent difficulties for analyzing the weights and bias terms [30].

Hidden Markov Model (HMM)

Algorithms based on HMMs are able to identify localization motifs by modeling sequence alignments. A test sequence is compared against all the HMM models produced for the various subcellular compartments, and it is assigned to the one with the highest probability score. The success of HMMs is highly dependent upon accurate alignment of the training sequences. Also, HMMs can only cope with linear sequences, and, for this reason, any localization motifs created as a result of the secondary or the tertiary structure cannot be modeled by HMMs. Modeling such 2D or 3D motifs may require alternative probabilistic methodologies such as stochastic context-free grammars (SCFGs). SCFGs have been successfully deployed to assist with the RNA secondary structure prediction [31] but have found limited application in predicting protein localization.

Covariant Discriminant Algorithm (CDA)

Subcellular localization predictors using CDAs appeared in the late 1990s [26–28]. Like most of their counterparts, they used the amino acid composition for their discriminatory analysis. Their predictive accuracy was soon outperformed by other approaches [19].

Apart from amino acid sequence and composition, additional information such as the GO annotation [32], sequence similarity using BLAST [33], text annotations from

protein sequence databases [25, 34, 35], or phylogeny [36] can be deployed in order to assist with the prediction process and boost the algorithm's performance.

Performance Comparison and Outlook

Due to the differences in the training and test data sets that have been used by the different localization predictions, it is difficult to compare their performances. To make matters more complicated, sequence similarity and localization identity are closely linked together [35], and, therefore, datasets containing highly similar or homologous sequences led to overestimated performance of the algorithms. Another obstacle for benchmarking the localization prediction algorithms is the difference in the number of cellular compartments that are predicted (Table 5.1). Thus the interpretation of results from benchmarking would not be straightforward even if using the same training and test datasets.

However, there has been a constant improvement of the localization prediction algorithms. Newly developed algorithms should be tailor-made for each species, taking into account the differences in the amino acid composition between organisms [15]. Also, there have been examples where hybrid algorithms performed better than each of the individuals, so it is likely that the new predictors will combine the best of each methodology. Thus, residents of compartments such as the nucleus can be predicted with relatively high accuracy. In contrast, members from other compartments like the vacuoles or peroxisomes are more difficult to predict. This is mainly due to poorly understood localization mechanisms and, consequently, to the lack of sufficient training data. Prediction accuracy for these less well characterized subcellular compartments is expected to rise as more relevant biological knowledge becomes available.

Bioinformatics-driven approaches alone, therefore, are unlikely to provide a comprehensive and detailed view of protein localization. Experimental approaches will continue to be necessary to reveal the true nature of protein localization and dynamics, particularly in specialized cell types and in response to different developmental and environmental signals.

5.2 EXPERIMENTAL DETERMINATION OF PROTEIN LOCALIZATION

This section reviews a variety of experimental methods that have been developed for determining protein localization and their contribution to understanding cellular structure

Cellular Fractionation and Biochemical Approaches

The first approaches aimed at defining subcellular localization of proteins depended on cell fractionation and later used immunolabeling techniques to confirm localization [37–39]. Cell fractionation protocols use the different physical properties of subcellular organelles to effect separation and partial purification. However, these approaches can be time-consuming, and cross-contamination with other subcellular organelles

is common. Additionally, only some compartments are amenable to fractionation. Finally, transient localization or localization restricted to small compartments [40; reviewed in references 41 and 42] is often missed. Despite these inherent disadvantages, the advent of proteomics has led to resurgence in the use of subcellular fractionation [reviewed in reference 42]. Different MS-based methods that allow identification of ever smaller amounts of protein, such as 2-DGE, MALDI, and ESI, use cell fractionation for the determination of protein localization [43; reviewed in references 44 and 45]. MS measurement of peptide fragment size has been used to determine the protein composition of yeast, the malaria parasite *P. falciparum*, and human cells [reviewed in reference 43]. Additionally, numerous complex subnuclear organelles have been characterized, including the human and plant nucleolus [13, 46, 47].

Many organelles are difficult to purify as a homogenous preparation, and cross-contamination presents a difficult problem. This is a particular problem for small membrane-bound organelles such as Golgi bodies, peroxisomes, and mitochondria. Dunkley and co-workers partially separated these organelles by density gradient centrifugation. Since the organelles are not completely separated, proteins from each organelle are present throughout the gradient; but by using cleavable isotopic tags and multivariate analysis of the resultant MS data, it was possible to match the distribution of unknown proteins with the distribution of proteins known to reside in the different organelles [48, 49]. Alternative approaches include free-flow electrophoresis [50].

Methods to Determine PPIs

Protein–Protein interactions are often essential for the correct subcellular localization of proteins, and misexpression of a given protein in the absence of its interacting partners can lead to mislocalization or degradation. There are various ways to examine protein interactions [reviewed in references 51 and 52]. Some methods require the lysis of cells and rely on protein interactions being stable enough to allow *in vitro* investigations: for example, immunoprecipitation, blue native-polyacrylamide gel electrophoresis (BN-PAGE), far-Western analysis, kinase assays, phage displays, and SPRC. *In vivo* techniques need a “reporter system” to indicate the presence of an interaction. These include Y2H interaction assays, protein fragment complementation assays, and FRET/BRET. Below we discuss some of the procedures utilized to study PPIs and indicate the limitations and advantages of each. It is good practice not to rely on the results from any one type of assay, but to perform a repertoire of supportive assays [53, 54]. One limiting factor of many biochemical-based assays is how to release the complexes from the different cell compartments and nonrelated proteins, while maintaining the integrity of those complexes.

Immunoprecipitation/Affinity Purification Methods

Immunoprecipitation relies on the affinity of an antibody attached to a solid phase (such as sepharose beads) to bind its antigenic protein in a complex, soluble protein mixture, thereby anchoring it to the sepharose. Nonspecifically bound proteins can be removed from the sepharose–antibody–protein complex by washing with buffers of

selected stringency. Proteins that interact with the antigenic protein can also be isolated in this manner, but controls essential to prove that the interaction is specific. At least two basic controls are required: incubating the protein mixture with (a) the solid phase alone (indicates if any proteins interact with the solid phase itself) and (b) an irrelevant antibody (preferably the pre-immune serum from the antibody host, which will reveal non-specific binding to serum proteins). Although immunoprecipitation is potentially very specific, the scale of isolation can be limited by a number of factors such as the abundance of the protein(s) to be isolated, the amount of antibody available (limiting in the case of a polyclonal antibody), and the yield required. Large-scale protein isolation requires the use of FPLC where the protein mixture passes through a column of solid phase. Small-scale isolations can yield results using just a few microliters of beads incubating in 10 volumes of cell extract, but this is sufficient when used with MS to identify interactions [55].

There are several limitations. First, the specificity of immunoprecipitation experiments relies on the quality of the antibody, the production of which can be very costly and time-consuming. Second, the intracellular localization of the interaction can be lost during the isolation and must be confirmed using independent methods [56]. Third, contamination from the antibody can obscure the target proteins, although it is possible to covalently cross-link the antibody to the solid phase (<http://www.piercenet.com/>) and elute the bound proteins with high salt or low pH buffers. This approach can work well with some antibodies but not others [57]. Another approach is to transform cells with a gene encoding an epitope-tagged protein that can be isolated using immobilized affinity chromatography that avoids antibodies. These include the tandem affinity purification (TAP) tag that utilizes a two-stage precipitation procedure [58]. First, the TAP-tagged protein is immunoprecipitated by virtue of an Ig-binding domain in the TAP sequence, the beads are washed as normal, and the protein is released from the beads by enzyme cleavage. The TAPa system [59] uses a 3C protease (at 4°C), whereas the TAPi system [60] uses TEV protease at 16°C. The free protein complex is then subjected to another round of purification using either IMAC (TAPa) or Calmodulin (CaM) affinity resin in the presence of calcium (TAPi) and finally eluted with buffers containing imidazole or EGTA, respectively.

Blue Native-Polyacrylamide Gel Electrophoresis (BN-PAGE)

Various techniques have been developed to reveal the composition of complexes, including BN-PAGE. BN-PAGE enables protein complexes to be separated under native conditions, and then the individual proteins from a particular complex are resolved under denaturing conditions. This technique was developed to examine solubilized membrane proteins from various tissues (mammalian muscle, yeast, and bacteria) and allowed separation of all the protein complexes of the oxidative phosphorylation system within one gel [61]. Their method was, until the publication of reference 62, not applicable for separating complexes from more complex mixtures such as whole-cell extracts. A simple dialysis step opened up the mammalian cell proteome to BN-PAGE separation and subsequent proteomic investigations.

BN-PAGE works by gently solubilizing cells to give a mixture of intact protein complexes, which can be separated on a non-denaturing gradient gel according to their

MW. The negative charge of the blue dye and the size and shape of each complex determines how far that complex migrates into the gel. The gel lane is then cut out and separated on a second gel orientated perpendicular to the first axis of separation and includes SDS to separate the component proteins of each complex according to their MW. BN-PAGE studies so far in plants have been restricted to membrane proteins [reviewed in reference 63]. Drykova et al. [64] used several techniques including BN-PAGE, to confirm that γ -tubulin is associated with membranes in the form of large (>1 MDa) protein complexes. Respiratory enzyme complexes in mitochondria and protein-import complexes in chloroplasts have been studied in *Arabidopsis* in references 63 and 65, respectively.

However, not all protein complexes appear to be suitably solubilized by any one detergent, and BN-PAGE does not resolve small protein complexes (<100 kDa) well due to the small separation distance of the first gel step. Distinct complexes of similar molecular masses may co-migrate, and the constitutive proteins then appear to be present in the same complex.

Far-Western Blotting

Originally developed to screen protein expression libraries [66], far-Western blotting (or overlay assay) is now used to identify and confirm PPIs [67]. In a far-Western study, the protein sample (e.g., a cell extract containing the unknown interactor or “prey”) is separated by SDS- or native PAGE and then transferred to a membrane. The membrane is blocked to prevent nonspecific binding and then probed with a known “bait” protein, which either is radiolabeled, bears a tag, or is identifiable with a specific antibody. The detection method is usually either direct (radiolabeled “bait”) or indirect (with an antibody to the “bait” protein or to the tag). Each method has advantages and disadvantages. The isotopes ^{35}S -Met [68] and ^{32}P are commonly used to label “baits,” the latter at phosphorylation sites within the fusion tags because this has little effect on the protein interaction. Glutathione *S*-transferase (GST)- or polyhistidine-tagged “bait” proteins are detected with a tag-specific antibody conjugated to HRP or AP [69]. If recombinant tagged proteins and bait-specific antibodies cannot be generated, then bait proteins can be biotinylated and immunolabeled with streptavidin antibodies.

5.3 IN VIVO IMAGING APPROACHES TO PROTEIN LOCALIZATION AND INTERACTION

Since fractionation necessitates cell breakage, there are obvious concerns regarding contamination and other artifacts that would lead to incorrect conclusions about localization. Therefore, independent verification of subcellular localization is normally required. In the past, verification depended largely on immunolabeling of whole cells. This method provides good spatial resolution as well as high specificity, but the requirement for specific antibodies can be expensive and time-consuming. Immunolocalization also requires fixation, combined with either tissue sectioning or digestion of

the cell wall to allow access of the antibody to the antigen [70]. A major disadvantage of both cell fractionation and immunolocalization is the requirement to “kill” the cell, because this makes the study of protein dynamics difficult.

To overcome these problems, a variety of real-time imaging approaches have been developed to follow proteins within living cells. Dynamic protein trafficking and localization was first revealed using microinjection to introduce fluorescently labeled antibodies and proteins into living cells. Specific proteins could then be visualized under the microscope, but this method was time-consuming and was generally restricted to those proteins that could be isolated in sufficient quantity and purity for *in vitro* labeling. Therefore, this technique is generally used only in favored cell types and is an extremely skilled operation [71, 72].

As gene cloning technologies were developed, it was found that genes of interest could be fused to reporter genes, such as *gus* [73] and *luc* and the fusion proteins detected, thereby revealing their subcellular location. These reporters are not ideal for high-resolution cell biology: Although GUS fusion proteins can be used to distinguish between nuclear and non-nuclear localization, detection depends on a colored precipitate that is prone to diffusion and so the spatial resolution is rather low. Luc assays require specific equipment for microscopic observation (sensitive photon counting cameras) and can suffer from high nonspecific background signal. In both cases, visualization requires addition of exogenous substrates and co-factors [74–78].

In the late 1980s, Chalfie’s lab developed green fluorescent protein (GFP), from the jellyfish *A. victoria*, as a generally applicable method for tagging almost any protein and following its location in living cells. GFP allows detection of protein expression without requiring any substrates or co-factors [74]. Also, fluorescence is detected in a noninvasive way and, for many tissues and plant species, is spectrally distinct from the autofluorescent pigments found within the cell [76, 79]. These properties of GFP established the protein as an invaluable tool for many fields of biological research [80, 81]. With minor modifications, this approach has been applied to a wide range of species. Translational fusions between the protein of interest, and GFPs are now routinely used for determining subcellular localization, dynamics, and turnover in bacteria [82], yeast [83, 84], mammalian cells, and plants [85, 86]. The utility of GFP is illustrated by the increasing number of GFP-related references. In a literature search with GFP in the title and/or abstract, 10 publications were found in 1994, while in 2002 over 6000 articles were indexed in PubMed [87]. At present, the number of GFP-related articles indexed in PubMed has reached 12,283 (searched in 2/5/2007). GFP also lends itself to high-throughput screens. Several model systems, such as yeast, *Arabidopsis*, and HeLa cells, have been used for screening cDNA libraries and full-length gene products, fused to GFP [6, 8, 9, 11, 13, 14, 88–91].

Apart from the GFP of *A. victoria*, several GFP-like proteins have been found in members of the Anthozoa class (coral animals). These proteins are responsible for most of the colors found in reef Anthozoa, where each color is being determined by a single protein. So far, about 100 GFP-like proteins have been cloned that display similar structural, biochemical, and photophysical properties with GFP but also show spectral and chromophore diversity [92–97]. Fluorescent proteins of various colours have been used to simultaneously visualize two or three proteins within a cell [98, 99].

Such fluorescent markers revolutionized localization studies and allowed observation of organelle biogenesis and dynamics, as well as examination of protein transport within and between living cells.

Transformation Methodologies for Studying Protein Localization

A key technology required for high-throughput protein localization is the ability to transform cells with high efficiency. After the gene of interest is fused to a marker, it has to be delivered and expressed within a cell or tissue that undergoes the biological function of interest. Although cell suspensions suffice for housekeeping functions such as cell division, a differentiated cell where appropriate organelles and compartments are clearly visible may be more advantageous [89, 90]. Within this context, we refer here to several transformation methodologies that could be considered for this purpose.

Protoplast Transformation with Naked DNA. Protoplast transformation has been used successfully for investigating a broad spectrum of plant signaling systems, for performing functional genomic and proteomic analysis, and for studying PPIs [100–103]. However, protoplasts generally do not divide and longer-term protein dynamics can be difficult to follow. Also, maintenance of the transformed cultures is labor-intensive and time-consuming because they require sterile conditions [104; and reviewed in reference 105].

Particle Bombardment. Particle bombardment (biolistics) is another “physical” transformation methodology that can target cells within plants and thus minimize the laborious and time-consuming procedures of tissue culture. However, this system is most suitable for targeting big vacuolated cells that generally do not divide [106, 107].

Virus Mediated-Transformation. Virus-mediated protein expression [90] exploits the ability of viruses to replicate in high levels inside cells. Protein expression is detectable within 3–4 days. Viral vectors, including *tobacco mosaic virus* (TMV) and *potato virus X* (PVX), have become an attractive alternate to plant transformation, mainly due to their simplicity and amenability to high-throughput screens [90]. However, the viral genomes are intolerant of large insertions, and this limits the size of fusion proteins that can be produced [108–110].

Agrobacterium-Mediated Transformation. Currently, the most widely used method for cell transformation depends on the natural genetic engineering ability of the plant pathogen *A. tumefaciens*. This bacterium was identified as the causative agent of crown gall, a disease that is characterized by the formation of tumors at the infected areas of plant tissue [111]. Since then, a number of studies have been conducted to discover the fundamental aspects of *Agrobacterium*–plant interactions. The molecular mechanisms of the plant transformation process were first elucidated in mid-1970s when it was suggested that fragments of oncogenetic DNA are required for bacterial infection of plant cells [112]. *A. tumefaciens* is able to transform many different eukaryotic species such as yeast [113], several species of filamentous fungi

[114], human cells [115], and plants [116] and represents a very significant tool for molecular biology. Recently, bacterial species outside the *Agrobacterium* genus such as *Rhizobium*, *Sinorhizobium*, and *Mesorhizobium* strains have also been used for plant transformation [117], but *Agrobacterium* strains remain the standard for most plant species. It should be noted, however, that many different *Agrobacterium* strains exist and they vary in their ability to transform particular plant species and cell types, indicating that an appropriate match between the bacterial strain and the genotype of the host is required for optimal transformation [118].

Generating GFP–Protein Fusions for Localization Assays

Fusion of a reporter protein, such as GFP, to any given protein can perturb its localization and/or function and several factors must be considered when generating GFP fusions. In particular, N- and C-terminal location signals can be occluded and it is common practice to produce at least two versions for each protein, one where GFP is added to the N-terminus and one where it is added to the C-terminus. In many cases, both versions produce the same pattern but a significant number do not. For example, mitochondrial proteins and proteins of the secretory pathway [6] tend to have terminal localization signals that are perturbed by proximal GFB fusions. In a few cases, both N- and C-terminal fusions do not provide a faithful and functional reporter and GFP must be inserted internally. This is exemplified by the auxin efflux carrier PIN1 and the glycosylphosphatidylinositol (GPI)-anchored protein Sku5 [88]. For both proteins, the N- and C-terminal fusion products were not functional but the internal GFP fusions were functional. In general, the barrel-like structure of GFP protects its overall conformation and ensures that attachment with a host protein will not eliminate fluorescence activity [80]. However, negative results are rarely published and the fluorescence of GFB fusion proteins do vary although in some cases this may be due to protein turnover and degradation. Ideally, all fusion proteins should be functionally tested before extensive use, but this has not been extensively applied. Large scale analysis of protein localization in yeast resulted in discrepancies between the various studies [91, 119]. The diversity in localization patterns described for specific proteins was attributed to the different type of GFP-fusion used on each occasion.

Overexpression of GFP–Protein Fusions

Many studies lead to overexpression of the fusion proteins (relative to the levels of the wild-type proteins) because strong constitutive promoters are used to drive gene expression. Constitutive promoters generally provide higher fluorescent intensity and make the determination of protein location easier, but it should be recognized that possible artifacts do occur, resulting in misinterpretation of the localization data [82]. Attempts to solve this problem have met with limited success in plants, in large part due to the absence of an effective site-specific recombination system that would allow precise genomic engineering to place the GFP fusion under the correct expression controls. Fluorescent tagging of full-length proteins (FTFLP) [88] uses internal tagging

at selected sites within genes and aims to retain the native gene regulatory sequences. However, random integration of these constructs can place them in different genomic contexts that could frustrate the original intention, and the complexity the constructions has limited their use.

5.4 PLANT CELL CULTURES FOR STUDYING PROTEIN LOCALIZATION

In parallel with the development of fluorescent markers and the advances in DNA transfer technology, cell cultures have been exploited as model systems for studying protein localization. Animal model systems, such as the human HeLa cell line, have been used successfully for performing systematic, *in vivo*, subcellular localization assays [6, 9, 91]. The use of plant cell suspensions for determining protein location revolutionized plant subcellular localization studies, because they allowed visualization of protein expression independently of tissues and organs. This is very important in localization assays using fluorescent markers, where many plant tissues produce autofluorescent pigments that make detection of GFP difficult.

The tobacco BY2 system has been one of the most successful examples for studying intracellular protein localization in plants cell synchronization. This cell culture has been widely used as a model for cytoskeleton studies [120–123] and for cell cycle research [124–127]. Additionally, the relatively large size of BY2 cells is appropriate for microscopic analysis of Golgi dynamics and for observation of membrane trafficking [128–131]. On the other hand, *Arabidopsis* cell suspensions have not been used widely in cell biology studies, mainly because of their small size, which is not ideal for observing intracellular organization and content. Also, *Arabidopsis* cell suspensions were thought to be problematic for cell cycle studies. However, recent successful synchronisation and cell cycle gene expression studies using *Arabidopsis* cell cultures indicate their value in cell cycle research [132–135].

The *Arabidopsis* Transient System

Despite the various existing transformation methodologies, a simple and efficient method that can be used for rapid high-throughput functional analysis of plant genes is required. Recently, we developed an *Agrobacterium*-mediated transient expression system for *Arabidopsis* (Columbia-0) cell suspensions that can be used for *in vivo* systematic protein localization assays [136]. The *Arabidopsis* transient system was based on an intron-tagged epitope-labeling assay used to study protein interactions *in vivo* [137–139]. We found that this cell line was also ideal to study protein localization and gives data comparable to that of stable transformation in tobacco BY2 cells (Figure 5.1), but with the advantage that producing and maintaining discrete transgenic cell lines are not required. Finally, individual repetitions of transient assays with a training set of proteins suggested that this system is highly reproducible and reliable and can be used for protein localization screens (Figure 5.2; also see Table 5.2).

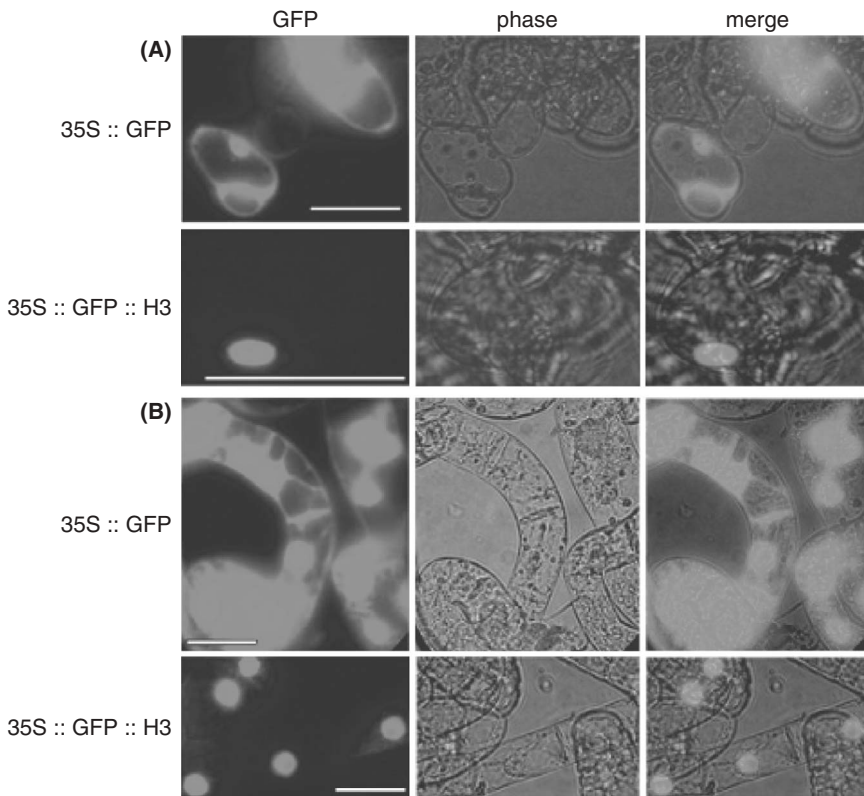


FIGURE 5.1. Comparison of GFP fusion protein localization in *Arabidopsis* culture cells (A) and tobacco BY2 cells (B). *Arabidopsis* cells have been transiently transformed, while BY2 cells were transformed and selected to produce stable cell lines. Patterns for GFP (unfused) and GFP-histone are shown. Bar represents 20 μm . See insert for color representation of this figure.

The *Arabidopsis* transient system has been used to determine successfully the sub-cellular location and follow the dynamic localization of diverse proteins [136]. The method has been applied to numerous cell biology studies in the lab [140–142] and as well as for high-throughput assays aimed at validating proteomic studies [11, 13]. The transient system is also versatile: Mutant dominant loss-of-function mutant constructs can be assayed because cells do not have to survive long term while co-transformation of multiple constructs allows co-localization of presumptive interacting proteins [156] and should permit FRET and other interaction assays.

5.5 PROTEIN–PROTEIN INTERACTION *IN VIVO*: FRET

Stryer [157] proposed that FRET could be used to probe close proximity interactions using two different fluorophores. One donates energy (the donor) to the other (the acceptor) when they are less than 10 nm apart; so if donor and acceptor are attached

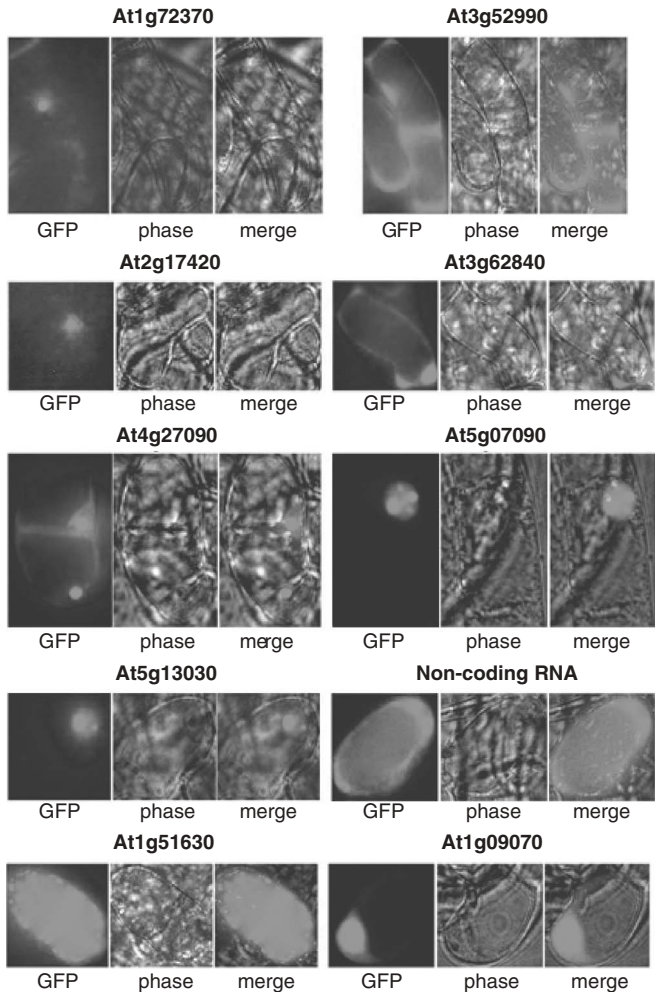


FIGURE 5.2. GFP-fusion localization showing a selection of patterns from a GFP localization screen based on transient transformation of *Arabidopsis* cells. Gene names are shown below the images and described in Table 5.2. (Bar represents 10 μm , except in At5g07090 where bar represents 7 μm). See insert for color representation of this figure

to two interacting proteins, the change in wavelength of the emitted fluorescence is indicative of an interaction. FRET relies on the donor and acceptor fluorophores having overlapping emission and excitation spectra, respectively, and can be measured in different ways. The change in emission intensities (donor quenching whilst the acceptor increases) are detected by fluorescence spectral microscopy (FSPIM), whereas the kinetics of the decaying donor emission are determined by fluorescence lifetime imaging microscopy (FLIM). Cyan fluorescent protein (CFP) as the donor and YFP as the acceptor are commonly used fluorophores combinations [143–146].

TABLE 5.2. Description of Genes for which Localization Patterns for Respective GFP Fusions are Shown in Figure 5.2

Atg number	Description	Biological Process
At1g72370	40S ribosomal protein SA	Mature ribosome assembly
At3g52990	Pyruvate kinase, putative	Involved in glycolysis
At2g17420	Thioredoxin reductase 2 (NADPH-dependent thioredoxin protein)	Thioredoxin reductase (NADPH) activity
At3g62840	Small nuclear ribonucleoprotein-related protein	Unknown
At4g27090	60S ribosomal protein L14	Protein biosynthesis
At5g07090	40S ribosomal protein S4	Protein biosynthesis
At5g13030	Expressed protein	Unknown
—	Noncoding RNA	Unknown
At1g51630	Expressed protein	Unknown
At1g09070	C2-domain containing protein	G-protein-coupled receptor protein signaling pathway

BRET, developed by Xu et al. [147], avoids the inherent problems with the FRET system: relatively high levels of cellular autofluorescence, the problem of photobleaching, direct excitation of the acceptor, and damage that could be done to very delicate tissues. BRET uses the blue light emitter *Renilla* luciferase as the donor and uses GFP or YFP as the acceptors of the bioluminescence energy. Subramanian et al. [148] adapted BRET for use in living onion epidermal cells and used *Arabidopsis* seedling roots to study COP1 protein interactions and subsequently went on to optimize BRET for *Arabidopsis* and *Nicotiana*.

Protein Fragment Complementation Assays (PCA)

These assays are based upon the principle that a polypeptide can fold itself into its active configuration and, that if the polypeptide chain is split into two parts, these separate peptides can refold into an active conformation when they are brought back together. If the peptide parts are derived from a reporter protein of some sort and are cloned one onto a protein “bait” and the other onto a “prey,” then when bait and prey interact the reporter protein can reform and be detected. PCA can be carried out in a cell type of choice, is independent of yeast and not restricted to the nucleus, so it allows interactions between proteins that would not be suitable for Y2H studies. The early PCAs used ubiquitin (Ub) split into inactive N- and C-terminal fragments, and the reformed active protein cleaves off a TF that enters the nucleus to activate a reporter gene [149]. This was an indirect system, so fluorophores such as GFP and YFP have been split and cloned onto the interacting proteins, a technique known as bimolecular fluorescence complementation (BiFC). This has the advantage of being able to view the protein interaction in living cells over extended periods without the need to add chromogenic substrates that might affect the cell physiology. The limitations are that the refolding of the split proteins occurs on the order of several

minutes, is thought to be irreversible *in vitro* (hence probably also *in vivo*), and may therefore be not suitable for highly dynamic protein interactions. Despite this, BiFC has been used to investigate *in vivo* protein interactions in *Arabidopsis* and *Nicotiana* leaves and protoplasts [150–153].

5.6 PERSPECTIVES: INTEGRATING PREDICTIVE AND EXPERIMENTAL PROTEIN LOCALIZATION DATA

Our knowledge of protein localization in plant cells is increasing rapidly. Because recent technological developments have made characterization of protein localization and dynamics so much more straightforward, it is now expected that acquiring these data is a requirement for any serious analysis of gene function. This has resulted in at least 1000 different protein localization being described over the past few years in *Arabidopsis*. However, there is now an urgent need for new forms of information tracking systems to catalogue this new image-based data and relate it to other forms of biological data. Heazlewood and co-workers have begun this task with the creation of a database that combines localization and proteomic data with predictive tools and links to the relevant literature and other databases (<http://www.suba.bcs.uwa.edu.au>). It is hoped that over the next few years this will lead to the population of genomic databases with high-quality accurate information on protein localization.

REFERENCES

1. Sheer U and Hock R. 1999. *Curr Opin Cell Biol* **11**:385–390.
2. Shaw P and Doonan J. 2005. *Cell Cycle* **4**:102–105.
3. Wasteneys GO and Galway ME. 2003. *Annu Rev Plant Biol* **54**:691–722.
4. Blobel G, et al. 1979. *Symp Soc Exp Biol* **33**:9–36.
5. Eisenhaber F and Bork P. 1998. *Trends Cell Biol* **8**:169–170.
6. Simpson JC, et al. 2000. *EMBO Rep* **1**:287–292.
7. Burns N, et al. 1994. *Genes Dev* **8**:1087–1105.
8. Cutler SR, et al. 2000. *Proc Natl Acad Sci USA* **97**:3718–3723.
9. Ding DQ, et al. 2000. *Genes Cells* **5**:169–190.
10. Kleffmann T, et al. 2004. *Curr Biol* **14**:354–362.
11. Koroleva OA, et al. 2005. *Plant J* **41**:162–174.
12. Kumar A, et al. 2002. *Genes Dev* **16**:707–719.
13. Pendle AF, et al. 2005. *Mol Biol Cell* **16**:260–269.
14. Sawin KE and Nurse P. 1996. *Proc Natl Acad Sci USA* **93**:15146–15151.
15. Schneider G and Fechner U. 2004. *Proteomics* **4**:1571–1580.
16. Nakai K and Horton P. 1999. *Trends Biochem Sci* **24**:34–36.
17. Horton P, et al. 2007. *Nucleic Acids Res* **35**:W585–W587.
18. Sarda D, et al. 2005. *BMC Bioinformatics* **6**:152.
19. Park KJ and Kanehisa M. 2003. *Bioinformatics* **19**:1656–1663.

20. Nair R and Rost B. 2005. *J Mol Biol* **348**:85–100.
21. Guo J, et al. 2006. *J Bioinform Comput Biol* **4**:1181–1195.
22. Small I, et al. 2004. *Proteomics* **4**:1581–1590.
23. Emanuelsson O, Nielsen H and von Heijne G. 1999. *Protein Sci* **8**:978–984.
24. Emanuelsson O, et al. 2000. *J Mol Biol* **300**:1005–1016.
25. Lu Z, et al. 2004. *Bioinformatics* **20**:547–556.
26. Chou KC and Elrod DW. 1999. *Protein Eng* **12**:107–118.
27. Chou KC and Elrod DW. 1998. *Biochem Biophys Res Commun* **252**:63–68.
28. Cedano J, et al. 1997. *J Mol Biol* **266**:594–600.
29. Schneider G and Wrede P. 1998. *Prog Biophys Mol Biol* **70**:175–222.
30. Almeida JS. 2002. *Curr Opin Biotechnol* **13**:72–76.
31. Do CB, Woods DA, and Batzoglou S. 2006. *Bioinformatics* **22**:e90–e98.
32. Chou KC and Cai YD. 2003. *Biochem Biophys Res Commun* **311**:743–747.
33. Pierleoni A, et al. 2006. *Bioinformatics* **22**:e408–e416.
34. Scott MS, Thomas DY, and Hallett MT. 2004. *Genome Res* **14**:1957–1966.
35. Nair R and Rost B. 2002. *Bioinformatics* **18**:S78–S86.
36. Marcotte EM, et al. 2000. *Proc Natl Acad Sci USA* **97**:12115–12120.
37. Singer SJ. 1959. *Nature* **183**:1523–1524.
38. Faulk WP and Taylor GM. 1971. *Immunochemistry* **8**:1081–1083.
39. Price CA. 1974. *Methods Enzymol* **31**:501–519.
40. Gonzalez C and Bejarano LA. 2000. *Trends Cell Biol* **10**:162–165.
41. de Souza W and da Cunha-e-Silva NL. 2003. *Mem Inst Oswaldo Cruz* **98**:151–170.
42. Huber LA, Pfaller K, and Vietor I. 2003. *Circ Res* **92**:962–968.
43. Tyers M and Mann M. 2003. *Nature* **422**:193–197.
44. Aebersold R and Mann M. 2003. *Nature* **422**:198–207.
45. Resing KA and Ahn NG. 2005. *FEBS Lett* **579**:885–889.
46. Andersen JS, et al. 2002. *Curr Biol* **12**:1–11.
47. Scherl A, et al. 2002. *Mol Biol Cell* **13**:4100–4109.
48. Dunkley TP, et al. 2006. *Proc Natl Acad Sci USA* **103**:6518–6523.
49. Dunkley TP, et al. 2004. *Mol Cell Proteomics* **3**:1128–1134.
50. Bardy N, et al. 1998. *Electrophoresis* **19**:1145–1153.
51. Kersten B, et al. 2006. *Proteomics* **6**:5517–5528.
52. Monti M, et al. 2005. *Biosci Rep* **25**:45–56.
53. Hazbun TR, et al. 2003. *Mol Cell* **12**:1353–1365.
54. Sang Y, et al. 2005. *Plant Cell* **17**:1569–1584.
55. Hutchins AP, et al. 2004. *FEBS Lett* **556**:91–94.
56. Cristea IM, et al. 2005. *Mol Cell Proteomics* **4**:1933–1941.
57. Delgado IJ, et al. 1998. *Plant Physiol* **116**:1339–1350.
58. Brown AP, et al. 2006. *J Exp Bot* **57**:1563–1571.
59. Rubio V, et al. 2005. *Plant J* **41**:767–778.
60. Rohila JS, et al. 2004. *Plant J* **38**:172–181.
61. Schagger H and von Jagow G. 1991. *Anal Biochem* **199**:223–231.

62. Camacho-Carvajal MM, et al. 2004. *Mol Cell Proteomics* **3**:176–182.
63. Dudkina NV, et al. 2005. *Proc Natl Acad Sci USA* **102**:3225–3229.
64. Drykova D, et al. 2003. *Plant Cell* **15**:465–480.
65. Chen KY and Li HM. 2007. *Plant J* **49**:149–158.
66. Blackwood EM and Eisenman RN. 1991. *Science* **251**:1211–1217.
67. Hall RA. 2004. *Methods Mol Biol* **261**:167–174.
68. Strathmann A, et al. 2001. *Plant J* **28**:397–408.
69. Yambao ML, et al. 2003. *J Gen Virol* **84**:2861–2869.
70. von Arnim AG, Deng XW, and Stacey MG. 1998. *Gene* **221**:35–43.
71. Celis JE, Graessmann A, and Loyter A. 1986. *Microinjection and Organelle Transplantation Techniques: Methods and Applications*, Academic Press (London) Ltd., London.
72. Lamb NJ, Gauthier-Rouviere C, and Fernandez A. 1996. *Front Biosci* **1**:d19–d29.
73. Jefferson RA, Burgess SM, and Hirsh D. 1986. *Proc Natl Acad Sci USA* **83**:8447–8451.
74. Chalfie M, et al. 1994. *Science* **263**:802–805.
75. Hu W and Cheng CL 1995. *FEBS Lett* **369**:331–334.
76. Haseloff J, et al. 1997. *Proc Natl Acad Sci USA* **94**:2122–2127.
77. Brandizzi F, Fricker M and Hawes C. 2002. *Nat Rev Mol Cell Biol* **3**:520–530.
78. Greer LF, 3rd and Szalay AA. 2002. *Luminescence* **17**:43–74.
79. Galbraith DW, Anderson MT, and Herzenberg LA. 1999. *Methods Cell Biol* **58**:315–341.
80. Misteli T and Spector DL. 1997. *Nat Biotechnol* **15**:961–964.
81. Prasher DC. 1995. *Trends Genet* **11**:320–323.
82. Margolin W. 2000. *Methods* **20**:62–72.
83. Ueki N, et al. 1998. *Nat Biotechnol* **16**:1338–1342.
84. Kohlwein SD. 2000. *Microsc Res Tech* **51**:511–529.
85. Haseloff J and Siemering KR. 1998. The uses of GFP in plants. In: *Green Fluorescent Protein: properties, Applications and Protocols* (Chalfie M and Kain S, eds.), Wiley, Hoboken, NJ, pp. 191–220.
86. Kohler RH. 1998. *Trends Plant Sci* **3**:317–320.
87. Zimmer M. 2002. *Chem Rev* **102**:759–781.
88. Tian GW, et al. 2004. *Plant Physiol* **135**:25–38.
89. Misawa K, et al. 2000. *Proc Natl Acad Sci USA* **97**:3062–3066.
90. Escobar NM, et al. 2003. *Plant Cell* **15**:1507–1523.
91. Huh WK, et al. 2003. *Nature* **425**:686–691.
92. Verkhusha VV and Lukyanov KA. 2004. *Nat Biotechnol* **22**:289–296.
93. Matz MV, et al. 1999. *Nat Biotechnol* **17**:969–973.
94. Lukyanov KA, et al. 2000. *J Biol Chem* **275**:25879–25882.
95. Wiedenmann J, et al. 2000. *Proc Natl Acad Sci USA* **97**:14091–14096.
96. Gurskaya NG, et al. 2001. *FEBS Lett* **507**:16–20.
97. Labas YA, et al. 2002. *Proc Natl Acad Sci USA* **99**:4256–4261.
98. Ellenberg J, Lippincott-Schwartz J, and Presley JF. 1999. *Trends Cell Biol* **9**:52–56.
99. Chan J, et al. 2005. *Plant Cell* **17**:1737–1748.
100. Cheng SH, et al. 2001. *FEBS Lett* **503**:185–188.

101. Sheen J. 2001. *Plant Physiol* **127**:1466–1475.
102. Subramaniam R, et al. 2001. *Nat Biotechnol* **19**:769–772.
103. Winfield S, et al. 2001. *In Vitro Cell Develop Biol—Plant* **37**:648–653.
104. Locatelli F, et al. 2003. *Plant Cell Rep* **21**:865–871.
105. Davey MR, et al. 2005. *Biotechnol Adv* **23**:131–171.
106. Christou P. 1996. *Trends Plant Sci* **1**:423–431.
107. Christou P. 1997. *Plant Mol Biol* **35**:197–203.
108. Takamatsu N, et al. 1987. *EMBO J* **6**:307–311.
109. Chapman S, Kavanagh T, and Baulcombe D. 1992. *Plant J* **2**:549–557.
110. Rommens CM, et al. 1995. *Plant Cell* **7**:249–257.
111. Smith EF and Townsend CO. 1907. *Science* **25**:671–673.
112. Chilton MD, et al. 1977. *Cell* **11**:263–271.
113. Piers KL, et al. 1996. *Proc Natl Acad Sci USA* **93**:1613–1618.
114. de Groot MJ, et al. 1998. *Nat Biotechnol* **16**:839–842.
115. Kunik T, et al. 2001. *Proc Natl Acad Sci USA* **98**:1871–1876.
116. Hooykaas PJ. 1989. *Plant Mol Biol* **13**:327–336.
117. Broothaerts W, et al. 2005. *Nature* **433**:629–633.
118. Clough SJ and Bent AF. 1998. *Plant J* **16**:735–743.
119. Ross-Macdonald P, et al. 1999. *Nature* **402**:413–418.
120. Hasezawa S and Kumagai F. 2002. *Int Rev Cytol* **214**:161–191.
121. Granger C and Cyr RJ. 2001. *J Cell Sci* **114**:599–607.
122. Granger CL and Cyr RJ. 2000. *Planta* **210**:502–509.
123. Dhonukshe P and Gadella TW Jr. 2003. *Plant Cell* **15**:597–611.
124. Criqui MC, et al. 2000. *Plant J* **24**:763–773.
125. Criqui MC, et al. 2001. *Plant J* **28**:569–581.
126. Weingartner M, et al. 2001. *Plant Cell* **13**:1929–1943.
127. Lee J, et al. 2003. *Plant J* **34**:417–425.
128. Nebenfuhr A, et al. 1999. *Plant Physiol* **121**:1127–1142.
129. Batoko H. et al. 2000 *Plant Cell* **12**:2201–2218.
130. Brandizzi F, et al. 2002. *Plant Cell* **14**:1293–1309.
131. Saint-Jore CM, et al. 2002. *Plant J* **29**:661–678.
132. Menges M, et al. 2003. *Plant Mol Biol* **53**:423–442.
133. Menges M and Murray JA. 2002. *Plant J* **30**:203–212.
134. Richard C, et al. 2001. *J Exp Bot* **52**:1625–1633.
135. Richard C, et al. 2002. *Plant Cell Tissue Organ Culture* **69**:167–176.
136. Kitsios G. 2006. Characterization of *Arabidopsis* cyclin-dependent kinases. Ph.D. Thesis, The University of East Anglia, Norwich.
137. Mathur J, et al. 1998. *Plant J* **13**:707–716.
138. Ferrando A, et al. 2000. *Plant J* **22**:553–560.
139. Ferrando A, et al. 2001. *Nucleic Acids Res* **29**:3685–3693.
140. Chan J, et al. 2003. *Nat Cell Biol* **5**:967–971.
141. Mao G, et al. 2005. *Plant J* **43**:469–478.

142. Korolev AV, et al. 2005. *Plant J* **42**:547–555.
143. Dhonukshe P, Bargmann BO, and Gadella TW Jr. 2006. *Plant Cell Physiol* **47**:1406–1411.
144. Feys BJ, et al. 2005. *Plant Cell* **17**:2601–2613.
145. Fu Y, et al. 2005. *Cell* **120**:687–700.
146. Huang Z, et al. 2001. *Plant Mol Biol* **47**:663–675.
147. Xu Y, Piston DW, and Johnson CH. 1999. *Proc Natl Acad Sci USA* **96**:151–156.
148. Subramanian C, et al. 2006. *Plant J* **48**:138–152.
149. Stagljar I, et al. 1998. *Proc Natl Acad Sci USA* **95**:5187–5192.
150. Bracha-Drori K, et al. 2004. *Plant J* **40**:419–427.
151. Jakoby MJ, et al. 2006. *Plant Physiol* **141**:1293–1305.
152. Liu X, et al. 2007. *Science* **315**:1712–1716.
153. Walter M, et al. 2004. *Plant J* **40**:428–438.
154. Bannai H, et al. 2002. *Bioinformatics* **18**:298–305.
155. Bodén M and Hawkins J. 2005. *Bioinformatics* **21**:2279–2286.
156. Kitsios G, et al. 2008. *Plant J* **54**:220–235.
157. Stryer 1978. *Annu Rev Biochem* **47**:819–846.

SECRETOME: TOWARD DECIPHERING THE SECRETORY PATHWAYS AND BEYOND

Young-Ho Jung, Ganesh Kumar Agrawal, Randeep Rakwal,
and Nam-Soo Jwa

6.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

The word “secretome” was first introduced in a global survey of secreted proteins of *B. subtilis* [1, 2]. Secretome describes the global study of secreted proteins by a cell, tissue, organ, or organism at any given time and conditions. Secretome is also a term referring to both the machineries for protein transport and the secreted proteins themselves. Although various transport systems have been studied in detail, little is known about temporal PPIs of secretome components and cellular or subcellular secretome localization. The secreted proteins constitute an important class of active molecules that control and regulate a multitude of biological and physiological processes in multicellular organisms, such as plants and animals. These proteins have been shown to act both locally and systemically [3–6]. Like other higher organisms, protein secretion in plants is a highly sophisticated and tightly regulated process. Despite decades of studies of the plant secretory pathways [reviewed in references 7–9], our knowledge on the secretory pathways and the mechanisms of protein secretion is highly limited. A list of proteins involved in the secretory pathways is still incomplete. Moreover,

little is known about the secreted proteins in plants in response to chemical elicitors or from pathogens upon plant–pathogen interaction.

Identification of a complete set of secreted proteins in reference plants like *Arabidopsis* and rice are required to begin deciphering the secretory pathways and the host–pathogen interactions on a global level. Proteomics approaches are particularly well-suited and preferred technologies for secretome analysis over gene expression approach, such as gene array-based transcriptomics study. This is because many secreted proteins are not expressed in abundance, are expressed by specialized cell types and during specific stages of development, or have an induced expression during specific cellular responses, including those in the innate immune response. Such compartmentalized secreted proteins are possible to isolate and enrich using techniques such as differential extraction. Use of proteomics in the case of plant–pathogen interactions is the key, as there is more than one genome present in this situation. Literature survey indicates that large-scale secretome studies have been conducted mostly in animals, yeast, and bacteria, but rarely in plants. *B. subtilis* and rat liver homogenates are good examples of a large-scale secretome analysis [1, 2, 10, 11].

The classical secretory pathway in plants is the endoplasmic reticulum (ER)/Golgi pathway, similar to that in the higher organisms such as animals. In this pathway, integral proteins of the plasma membrane (PM) as well as secreted proteins are typically synthesized at the ER and pass through the Golgi apparatus, although some secretory vesicles bypass the Golgi on their way to the vacuole [12]. The Golgi apparatus is a major sorting station where trafficking routes diverge to the PM, the endosome, the vacuole, or the cell plate [13], therefore, a tight cooperation and regulation among organelles are required for proper function of the ER/Golgi pathway. Hence, identification of protein components of this organelle is also important to better understand the plant secretory pathways. Part V in this book has solely been devoted to organelle proteomics that deals in detail about the plant organelle studies. We must emphasize that fewer efforts have been made toward understanding association of plant organelle with the ER/Golgi secretory pathway on a large scale. Recently, a comprehensive and quantitative proteomics analysis of the ER/Golgi secretory pathway was conducted in the rat liver homogenates isolated rough ER, smooth ER, and Golgi apparatus, where more than 1400 different proteins were assigned to the secretory pathway [14].

Despite the tremendous progress made in unraveling the plant organelle proteomes, the secretory pathways remain poorly understood. To avoid overlap with studies on plant organelle proteomes, this chapter mainly deals with the secreted proteins due to chemical elicitor or pathogen attack. There have been very limited proteomics studies conducted on secreted proteins in plants on a global level. Two studies are the secretome analysis of *Arabidopsis* cultured cells treated with chemical elicitors [15] and a comparative secretome analysis of extracellular proteins *in vitro* and *in planta* from pathogenic fungus [16].

Identification of secreted proteins in *Arabidopsis* cultured cells treated with and without salicylic acid (SA) was performed using 2D gel-based proteomics approach [15]. Based on the demonstrated role of SA in plant defense signaling [17], it was hypothesized that *Arabidopsis* cultured cells treated with and without SA will help in

identifying secreted proteins involved in the pathogen response. A silver nitrate-stained 2D gel analysis of untreated cultured cells revealed ~400 protein spots, mostly distributed in the 10- to 100-kDa range. A total of 107 protein spots were identified by MALDI-TOF-MS analysis, corresponding to 91 different proteins with variety of cellular functions. The largest functional class was the novel or uncharacterized proteins (34%), while proteins involved in signaling and regulation comprised the second largest functional class (29.6%).

Secreted proteins have been shown to bear a short stretch of amino acids called a signal peptide at their N-terminus [18]. The presence of a signal peptide in a protein is considered as a hallmark signature of secreted protein. SignalP software (<http://www.cbs.dtu.dk/services/SignalP>), which searches for signal sequences and their cleavages sites [18], is generally used to determine whether the identified secreted proteins contain signal peptides. SignalP analysis of secreted proteins identified by Oh et al. [15] revealed that only 54% of these proteins possessed signal peptides. This result suggests the presence of secretory pathway(s) alternative to classical ER/Golgi pathway. A similar observation/suggestion has been made in the secretome analysis of other organisms [1, 2, 10, 19, 20]. A comparative analysis of silver nitrate-stained 2D gels between untreated and SA (0.5 mM)-treated medium led to the identification of 13 different proteins whose protein spot abundance was altered relative to the untreated control medium. One of these proteins was identified as GDSL motif lipase/hydrolase-like protein, called GLIP1, and selected for its functional study. Functional analysis suggested that GLIP1 might be a critical component in plant resistance to *A. brassicicola*.

Walton's laboratory took 1-DGE- and shotgun-based proteomics approaches to systematically identify proteins secreted by *F. graminearum* (*G. zae*) during growth on 13 media *in vitro* and *in planta* during infection of wheat heads [16]. The culture filtrates and vacuum infiltration were used to collect *in vitro*- and *in planta*-secreted proteins. A total of 289 proteins (229 *in vitro* and 120 *in planta*) were identified. As expected, both approaches were found to be complementary. However, 1-DGE analysis yielded more proteins than any shotgun analysis mainly due to prefractionation of complex proteins and gave higher statistical reliability for the identified proteins. Comparison of *in vitro* and *in planta* secretome showed that the proteins identified *in planta* differed significantly from those found when *F. graminearum* is grown *in vitro*. Forty-nine of the *in planta* proteins were not found in any of the *in vitro* conditions. At least 13 "housekeeping" proteins predicted to be cytoplasmic were found only *in planta*. Whether or not the housekeeping and other proteins without signal peptides are actively secreted *in planta* or are in the apoplast, they might influence the course of the host-pathogen interaction by interacting with host cells and receptors. Interestingly, the majority of the *in vitro* proteins (more than 90%) were predicted to have signal peptides compared to around 50% of the *in planta* proteins, suggesting again the existence of nonclassical ER/Golgi secretory pathway in plants.

During the past few years, our laboratory also has been involved in systematic identification of secreted proteins in rice in response to chemical elicitors and pathogen infection. We are mainly interested in secreted proteins *in planta*. During the course of

this study, we have developed a workflow to prepare and enrich high-quality secreted proteins suitable for gel (2-DGE and 1-DGE) and non-gel (like MudPIT)-based proteomics analysis. We briefly discuss the unpublished methodology, strategy, and initial data on the identification of secreted proteins in the following sections.

6.2 METHODOLOGY AND STRATEGY

Secretome analysis in plants is only beginning to emerge. A survey of limited studies shows that techniques for developing a highly confident secretome of plant cell/tissue have yet to be fully standardized to ensure reproducibility from laboratory to laboratory. It is therefore important to establish a workflow for systematic secretome analysis of plant species. A general consideration to investigate plant secretome on a large scale should be: isolation of pure and intact secreted proteins, application of complementary proteomics approaches, development of high-resolution 2D reference map, high-quality MS and MS/MS spectra generation to assign peptide sequence with high accuracy, quantitative data on identified proteins, and database development and dissemination to the scientific community.

In vitro-cultured cell systems have widely been used to prepare secreted proteins. Though an *in vitro* system is easy to maintain and handle, the system does not represent the physiological processes of plants. Therefore, investigation of secretome *in planta* is essential. In the case of cultured cells, the culture filtrates are used for extraction of secreted proteins, whereas vacuum infiltration method has widely been used to extract secreted proteins *in planta*. The culture filtrates or vacuum infiltrates are centrifuged to obtain clear supernatant. The clear supernatant is generally freeze-dried in a vacuum lyophilizer, and the pellet resuspended in small volume of water (approximately 2 mL) is either dialyzed against suitable buffer like MES (10 mM, pH 5.8) or passed through a desalting column (for example, PD-10 with 5-kDa MW cutoff from Bio-Rad), followed again by vacuum lyophilization. The secreted proteins pellet thus prepared can be stored at -80°C until further use. The pellet has been used directly for 2-DGE studies [15] or subjected to TCA precipitation to remove other interfering impurities before conducting downstream analysis [16]. The prepared secreted proteins can be analyzed using 2-DGE- and MS-based proteomics approaches [15].

We found that the lyophilized pellet carries interfering impurities such as carbohydrates. These impurities should be removed prior to downstream proteomics analysis. We routinely use TCA and cetyltrimethylammonium bromide (CTAB) precipitation steps to remove interfering impurities (unpublished data; Jung et al., personal communication). CTAB precipitation seems to be necessary for secreted proteins derived from fungus-infected plants. Two-dimensional gels of protein samples prepared without CTAB step showed high background when stained with silver nitrate, masking low abundance proteins. High background might be due to carbohydrates in the protein sample. CTAB precipitation largely removes carbohydrates besides other interfering impurities. Therefore, we recommend performing TCA followed by CTAB precipitation steps to prepare *in planta* high-quality secreted proteins for secretome analysis.

All proteins identified using complementary proteomics approaches should be subjected for signal peptide analysis using SignalP prediction software [18, 21]. Prediction of transmembrane domains (TMDs) of proteins has also been used along with signal peptide prediction to predict the true secreted proteins [6]. These computational programs can be used to easily generate a comprehensive list of potentially secreted proteins in a given tissue or organism providing the availability of genome sequence and, if possible, gene expression data. However, proteomics-based (or other experimental) data are required to verify the computational generated data. In light of this view, *Arabidopsis* and rice are indeed excellent reference plants to begin investigating plant secretome.

6.3 A CASE STUDY: *IN PLANTA* AND *IN VITRO* PROTEIN PROFILES OF SOLUBLE AND SECRETED PROTEINS IN RICE

We present a case example of *in planta* and *in vitro* secretome analyses in rice. A 2D gel-based proteomics approach was applied to identify and quantify the secreted proteins (unpublished data; Jung et al., personal communication). As an example, selected areas of 2D profiles of *in planta* and *in vitro* experiments are shown in Figure 6.1. A comparative analysis of these 2D profiles indicates dramatic differences in the protein patterns of *in planta* and *in vitro* derived soluble and secreted proteins. Trypsin-digested protein spots were subjected to LC-MS/MS and database analyses to identify protein. Four protein spots marked in Figure 6.1 were identified as RuBisCO activase isoform precursor, glucan 1,3- β -glucosidase precursor, α -amylase isoform III, and aspartyl protease. Glucan 1,3- β -glucosidase precursor and α -amylase isoform III have previously been reported as secreted proteins. Inventory of all identified proteins reveals that secreted proteins are involved in a variety of cellular processes. Many novel and unknown proteins were also identified.

SignalP prediction analysis of all identified proteins indicated that secreted proteins possessing signal peptides were significantly higher in an *in vitro*-prepared secreted protein population compared to *in planta*. A similar result was found in a comparative analysis of *in planta*- and *in vitro*-secreted proteins [16]. Lack of signature signal peptide in approximately 50% identified *in planta*-secreted proteins along with studies conducted in independent laboratories on the global level supports the notion that plant possesses yet unidentified secretory pathway(s) in addition to the ER/Golgi pathway. Evidence in this direction is emerging. It has been shown that improperly folded GFP is secreted via a nonclassical pathway in the extracellular space of Chinese hamster ovary cells [22]. Availability of large-scale protein localization tools and resources of knockout mutants for *Arabidopsis* and rice will help in dissecting the secretory pathway by understanding the subcellular localization of an experimentally identified secretory protein lacking signal peptide.

6.4 CONCLUSIONS

Proteomics of secreted proteins in plants is at an initial stage, but has still produced meaningful data. These data are due to continuous improvement in techniques and

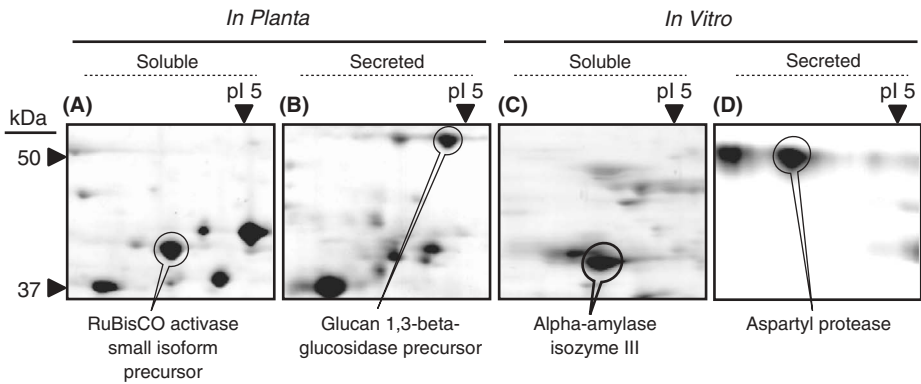


FIGURE 6.1. *In planta*- and *in vitro*-derived soluble and secreted proteins reveal dramatic differences in their 2D protein profiles. Both soluble and secreted proteins were isolated from 3-week-old rice seedling leaves treated with 250 ppm Tween 20 buffer for 24 h (called *in planta* experiment) and rice callus cultured in liquid medium (2 N6 [24, 25]) for 5 days after replacing the culture medium with the fresh culture medium (called *in vitro* experiment), respectively. Rice callus was generated from the rice seed growing on the induction medium [3% (w/v) sucrose, 0.03% (w/v) casamino acid, proline (2.878 mg/L), CHU (3.981 g/L), 1 mL of 2,4-D (2 mg/mL ethanol), pH 5.8, 0.2% (w/v) gellan gum] for three weeks. Growing calli (0.5 to 1 g) were transferred into 50 mL induction liquid media to establish cell suspension culture. The cell suspension culture was maintained by subculturing (1 mL) into 50 mL of fresh induction medium once every week with gentle agitation (120 rpm) at 28°C. Isolated proteins (150 µg) were separated on linear IPG strips (pH 4–7; 24 cm) in the first dimension followed by SDS-PAGE (12%) in the second dimension to develop 2D gel-secreted protein reference maps. Proteins were visualized with silver nitrate staining. The gel areas were selected from high-resolution 2D gels of an *in planta*- and *in vitro*-experiments to display the drastic differences in protein profiles between not only the soluble and secreted proteins but also between *in planta* and *in vitro* secreted proteins. The pI and molecular mass range (in kDa) of the selected gels are given. Soluble and secreted 2D protein profiles of *in planta* experiments are displayed in A and B, respectively, whereas *in vitro* experiments are displayed in C and D, respectively. Results presented here are representative of four independent biological experiments. Protein spots identified by LC–MS/MS and database analyses are also given for one protein spot per 2D gel. Except RuBisCO activase isoform precursor and aspartyl protease, glucan 1,3-β-glucosidase precursor and α-amylase isoform III have been previously reported as secreted proteins, suggesting that the applied workflow for preparation and enrichment of secreted proteins is suitable for investigating the secret of rice secretome.

workflow to properly investigate the plant secretome at a global scale. *In planta*- and *in vitro*-collected secreted proteins show dramatic differences in the 2D protein patterns and in the number of secreted proteins with a signal peptide, and therefore these experimental approaches can together be used to dig deeper into the plant secretome. Accumulated data have also begun providing insight on the nature of secreted proteins in plants under pathogen invasion or due to chemical elicitors. These proteins are of two kinds: proteins carrying and lacking signal peptides. Proteins lacking signal peptides imply the presence of nonclassical ER/Golgi secretory pathways in plants.

Lastly, based on studies conducted to date on plant secretome, though very limited, we must emphasize that proteomics approaches can be used to identify secreted proteins and to reveal their function.

6.5 FIVE-YEAR VIEWPOINT

Systematic investigation of plant secretome on a large scale has only recently begun and has to go a long way to further improve the existing technology and to develop a more suitable and reproducible workflow to identify all secreted proteins under normal and abnormal conditions. The list of secreted proteins has started to accumulate and is expected to grow at an unprecedented rate in the coming years. Identification of *in planta*-secreted proteins is specifically important, because they express under the given physiological conditions. This inventory will open a door for generating hypothesis and developing discovery-based projects to understand the role of identified secreted proteins. These experimentally identified secreted proteins are also expected to help in developing even better prediction programs for signal peptides that can eventually be used to design a more suitable computational strategy to predict *in silico*-secreted proteins and their function, such as physical properties and subcellular localization.

Arabidopsis resources have already been used to dissect the function of a secreted protein GLIP1 identified by the proteomics approach [15]. We envision that function of more secreted proteins, will be unraveled in the coming five years. Functional analysis of identified secreted proteins, notably unknown and novel proteins will help in getting insight into the secretory pathways other than the classical ER/Golgi pathway and into the mechanisms of protein secretion in plants. Studies conducted in a simple organism, *B. subtilis*, have shown the presence of at least four distinct secretory pathways for approximately 300 proteins predicted to be extracellular [1, 23].

Secreted proteins are also known to be involved in establishing the first contact between a pathogen and its host. Identification of secreted proteins will certainly increase our knowledge on plant–pathogen interactions. We do understand that this journey is quite challenging and sometimes frustrating due to the spectrum of involved multicellular processes and complexity, but at the same time fascinating because the term secretome encompasses both machineries for protein transport and secreted proteins. Decoding the secrets of secretome is the target of research for many ambitious biologists including us.

ACKNOWLEDGMENT

This work was supported by the faculty research fund of Sejong University in 2006.

REFERENCES

1. Tjalsma H, et al. 2000. *Microbiol Mol Biol Rev* **64**:515–547.
2. Antelmann H, et al. 2001. *Genome Res* **11**:1484–1502.
3. Oparka KJ and Cruz SS. 2000. *Annu Rev Plant Physiol Plant Mol Biol* **51**:323–347.

4. Oparka KJ. 2004. *Trends Plant Sci* **9**:33–41.
5. Pedersen BK and Febbraio M. 2005. *Brain Behav Immun* **19**:371–376.
6. Bortoluzzi S, et al. 2006. *Proteins* **62**:776–792.
7. Battey NH, et al. 1999. *Plant Cell* **11**:643–659.
8. Jurgens G and Geldner N. 2002. *Traffic* **3**:605–613.
9. Hanton SL, et al. 2007. *Plant Cell Rep* **26**:1431–1438.
10. Tjalsma H, et al. 2004. *Microbiol Mol Biol Rev* **68**:207–233.
11. Economou A. 2002. *Mol Membr Biol* **19**:159–169.
12. Dupree P. 1999. *Trends Cell Biol* **9**:130.
13. Sanderfoot AA and Raikhel NV. 1999. *Plant Cell* **11**:629–641.
14. Gilchrist A, et al. 2006. *Cell* **127**:1265–1281.
15. Oh S, et al. 2005. *Plant Cell* **17**:2832–2847.
16. Paper JM, et al. 2007. *Proteomics* **7**:3171–3183.
17. Durner J, Shah J, and Klessig DF. 1997. *Trends Plant Sci* **2**:266–274.
18. Nielsen H, et al. 1997. *Protein Eng* **10**:1–6.
19. Hathout Y. 2007. *Expert Rev Proteomics* **4**:239–248.
20. Zvonick S, et al. 2007. *Mol Cell Proteomics* **6**:18–28.
21. Bendtsen JD, et al. *J Mol Biol* **340**:783–795.
22. Tanudji M, Hevi S, and Chuck SL. 2002. *J Cell Sci* **115**:3849–3857.
23. van Dijl JM, et al. 2002. *J Biotechnol* **98**:243–254.
24. Chu CC. 1978. In: *Proceedings, Symposium on Plant Tissue Culture*, Science Press, Peking, pp. 43–50.
25. Hiei Y, et al. 1994. *Plant J* **6**:271–282.

PEPTIDOMICS

Peter Schulz-Knappe

7.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Peptides and proteins define the phenotype of all organisms, in form and in function. By utilizing a multitude of peptides and proteins, normal physiology and pathogenic processes are realized. This chapter will focus on a certain subpart of the proteome, the smaller native endogenous peptides. This novel research area and the technologies utilized in general are referred to as peptidomics, a merger of the term peptide proteomics. For a general review on peptidomics, we refer to reference 1. No general accepted definition separates peptides from proteins, so we use as working hypothesis the term peptide for native polypeptide chains up to the range of ~15 kDa, taking into account that others already refer to insulin (5.7 kDa) as a small protein. Currently there is little information available on peptidomics in plants. To cover this *terra incognita* of proteomics in this book, an overview shall be given about peptidomics in humans, and the basic rules and concepts are outlined which appear relevant to future peptidomics research in plants.

Several basic principles of biology contribute to a significant extent to the known complexity of higher organisms. It has to be noted that this complexity is mainly

generated on the basis of a definitive—and rather limited—number of genes. It appears that the increase in complexity in higher organisms is not achieved by simply increasing the number of genes compared to lower organisms, but by processes that address gene- and gene-product regulation and diversification. For example, the human genome consists of a total of 22,500 genes. This rather low number of genes is in striking contrast to the observed high level of biological diversity. As dominant drivers of diversity, alternative splicing and various forms of post-translational processing, such as proteolytic processing of proteins, are considered the most important principles to allow for time-resolved and space-restricted metabolic variation in living systems. In general, many ways of post-translational processing events have been described. They serve many intracellular as well as extracellular purposes and introduce dynamic modifications of the proteome.

After being introduced by us only a few years ago [2, 3] peptidomics (also called LMW proteomics) was applied in human diseases to search for biomarkers in the peptide region of the proteome. E. Petricoin and L. Liotta, at that time working at the National Cancer Institute, reported about the use of surface-enhanced laser desorption/ionization (SELDI)–TOF–MS for MS pattern acquisition from the LMW proteome in sera samples of patients with ovarian cancer [4]. By application of bioinformatics tools, they succeeded to separate cancer and control samples with excellent sensitivity and specificity, but without any information on the identity of the respective peptides. This approach has initiated much further work, and applications quickly moved to many fields, especially in cancer biomarker discovery, and to many different samples, utilizing both cell and tissue extracts as well as body fluids. Whereas in recent years several technological details of the initial reports were challenged, the concept of peptide biomarkers per se remains one of the most promising approaches in current medical science.

A general principle is applicable to proteins and peptides, which is their successive change in molecular forms over the lifetime of any gene product. A single protein will undergo processing that results in several molecular forms with defined, and sometimes quite different, biological roles. These molecular forms have to be considered as individual species that can occur in the same location at different concentrations, but also in different compartments of the organism, at the same time. Often they demonstrate a regulated temporal distribution. The description of a given proteome or peptidome can only be precise if we are able to assess the status of a highly complex mixture of thousands of peptides in relation to and dependent of location, concentration, and time.

Proteases and peptidases constitute the responsible class of biomolecules catalyzing the proteolytic hydrolysis reaction during the processing of proteins and peptides. In the human organism, more than 500 proteases are present. As antagonists, a large number of protease inhibitors are encoded in the genome, demonstrating a high degree of regulation and counter-regulation. Proteome and peptidome are closely linked one to another: The “input” into the peptidome equals the proteolytic “output” from the proteome. Major drivers are proteases and their inhibitors. If research in peptidomics is performed, proteases and protease inhibitors are always to be considered.

7.2 SEPARATION TECHNOLOGY

About 20 years ago the vision existed to utilize huge amounts of human blood to isolate and identify novel peptide hormones. This concept was realized in form of a peptide bank from human blood filtrate, and over the years more than 700,000 L of this blood filtrate were collected, subfractionated in native peptide fractions, and used in many assay systems [5]. It quickly became clear that at least 100,000 peptides are regularly present in human blood; and further work showed that also other sources are rich in peptides, with at least 5000 members per individual source. It is obvious that a combination of suitable technologies would be required to explore this large pool of peptides in living organisms. Given the relevance of bioactive peptides such as insulin, there was—and still is—quite some enthusiasm that the knowledge about native peptides is exceptionally useful in medical research. This view was supported by their involvement in key regulatory processes as outlined in Figure 7.1

A need for improved peptide discovery exists, ideally combining peptide sequence identification with quantitative peptide profiling, in order to find new peptides relevant as biomarkers and protopeptide drugs. The ideal system/workflow would fulfill several tasks, either subsequently or simultaneously: (i) It would be sensitive to an extent that relevant molecules can be detected, (ii) it would have a certain high degree of specificity to allow for unambiguous matching of analytes between samples, (iii) it would work in a (semi-) quantitative fashion to allow the comparison of all analytes between samples, (iv) it would deliver sequence information of the peptide of interest, and (v) it would utilize appropriate bioinformatics/biostatistics tools for multivariate and hypervariate statistics to deliver significant results with low FP percentages. If one compares these five tasks with the current development in proteomics, they are—not surprisingly—the same bullet points as have emerged in proteomics over the last years. We generally observe high dynamic excursions (i.e., biological variation is high), analyze hundreds to thousands of analytes simultaneously, face the presence of high abundant versus low abundant analytes, sometimes with a spread of over 10 orders of magnitude, and can only analyse a restricted number of samples. The risk to discover FPs is extremely high, and we have to interpret our own and published data with care, in order to deliver results of high quality to add relevant parts to current knowledge. The high number of analytes in combination with the biological variability, preanalytical variation during sample collection, and storage/processing and technical variation during analysis remains an extraordinary challenge. High-quality standards in reproducibility at all stages of the workflow, sufficient throughput (sample number, group size), and adequate dynamic range (quantification capabilities) of the analysis are a prerequisite for the performance of meaningful discovery studies.

Peptidomics technologies differ in several aspects from standard proteomics workflows. One standard approach, the 2-DGE, is not suitable for a systematic peptide analysis, since it does not cover peptides below 5–10 kDa. Other non-gel-based technologies such as MudPIT have become popular and are used in many variations. Common proteomics workflows make use of the protease trypsin, in order to digest proteins to make them more manageable by column chromatography and for MS analysis. We do not consider this as an appropriate way to analyze a peptidome. By

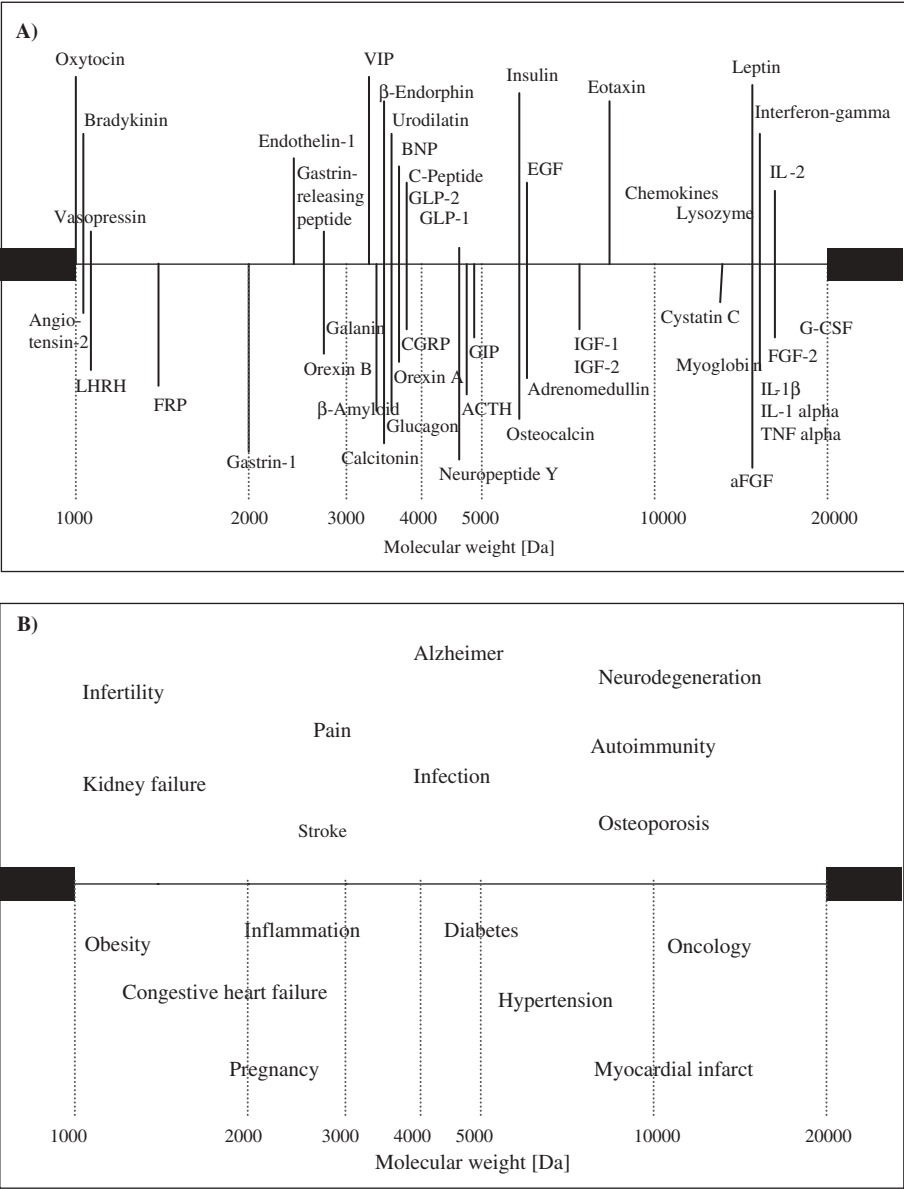


FIGURE 7.1. The peptidomics space (A, B).

Box 7.1**A list of key issues regarding peptidomics:**

- Peptidomics complements Proteomics for in-depth analysis of gene products in all organisms
- Peptides are especially important in extracellular environments
- Peptides are bioactive compounds
- Peptides are products of degeneration
- Peptides are ideal biomarkers
- Peptides mirror the activity of proteases
- The role of peptides in plants is still to be discovered

proteolysis of proteins, very high numbers and quantities of peptides are generated which have not been present in the sample itself. Therefore, the biological processing pattern of proteins to peptides is destroyed. Apart from limited sensitivity due to increased “protein background,” this makes the distinction between protein and peptide biomarkers impossible.

Usually in combination with (complete) protein removal, native peptides are extracted from samples and subsequently separated by means of chromatography. As good methods, we recommend a combination of cation-exchange chromatography (CAX) and reversed-phase (RP) chromatography. These chromatography methods show that excellent separation power can be performed in preparative to microanalytical scale and can be alternated in order to reduce sample complexity or enrich specific analytes. By using different pH values, buffer salts, and reversed-phase matrices, sufficient variation can be introduced to achieve very good separation even of very complex mixtures. If RP is the final separation step, the peptide sample should be in optimal condition for MS analysis.

7.3 MS TECHNOLOGY

In peptidomics, modern MS is an indispensable tool. For the purpose of sample discrimination, high sensitivity and specificity can be achieved during the discovery process by the analysis of individual peaks or by pattern algorithms alone. This has been extensively performed and published even by using rather simple mass spectrometers of the SELDI type. The caveat is that, without sequence information of the discriminating peptides, no in-depth information about biological mechanisms and the relevance of the discrimination is obtained. This will make any validation, reproduction of tests, and standardization for diagnostic applications a complicated task. A detailed sequence analysis will avoid the risk of focusing on peaks derived from medium to highly abundant molecules, which may be of more general biological relevance. Examples from human blood are acute-phase proteins, transport proteins, or other abundant proteins secreted from liver, which are frequently altered in many

disease states due to malnutrition, fever, hospitalization, and other circumstances and which are likely to be unspecific and of little scientific or commercial value.

MS is used for three general purposes: (i) qualitative analysis, (ii) quantitative analysis, and (iii) sequence analysis. First, a qualitative measurement, usually in MS mode, serves to find out how many different peptides are present in a sample by giving a certain number of (well enough) resolved peaks that follow a set of performance criteria (signal-to-noise ratio, peak shape, etc.). Second, the MS data are utilized for quantitative information on the individual analytes in the sample. This can be done in MS mode by comparing peak heights or peak integrals, and various concepts for alignment of masses, intensity normalization, and so on, are currently used. In a new set of applications, the quantification is performed by utilizing mass labels, either isotopic labels or isobaric labels. The clever concept is that two or more samples are differentially labeled early in the workflow and subsequently mixed. This helps to minimize technical variation in subsequent separation and purification steps (a similar approach is used in 2D gel-based proteomics by using fluorescent dyes, e.g., DIGE technology). When sample mixtures are then simultaneously analyzed in the mass spectrometer, isotopic labels will produce indicative peptide pairs in MS mode. Peak intensities are used for relative comparison of peptide concentration. Isobaric labels cannot be distinguished in single MS mode; they need to undergo fragmentation during MS/MS analysis. The different labels give rise to specific reporter ions, usually in a low-mass region of the mass spectrometer which is free of any amino acid derivatives. The relative abundance of the reporter ions is used for quantification. Since the peptides in peptidomics studies are not digested, they incorporate different numbers of labels, depending on the site where labeling occurs. Common reactive groups such as amines (in addition to N-termini the ϵ -amino-functions in lysins) will be increasingly frequent in native peptides, especially if they have longer polypeptide chains compared to tryptic digests. The mass spectrometrists (and data analyst) has to expect a variety of mass shifts in the case of isotopic labeling. This would make the selection of pairs quite challenging and error-prone, especially in the case of complex sample mixtures. Higher degrees of prefractionation may solve this problem. Third, the mass spectrometer is used to elucidate the amino acid sequence and possible PTMs of the peptides in a sample. In contrast to proteomics workflows, native peptides are usually analyzed in their intact form.

We wish to explain this in the following. In a typical sample, one can expect several peptides that derive from the same protein precursor. To illustrate this, we refer to the peptide insulin, which is a native cleavage product from the insulin precursor molecule. Normally, a combination of prohormone convertase cleavage and carboxypeptidase cleavage gives rise to the connecting peptide (C-peptide; 2.5 kDa) and the mature, bioactive insulin molecule (5.8 kDa), which has two polypeptide chains linked by two S–S bonds. In end-stage diabetes type II, this processing that takes place in the β -cells of the pancreatic islets is impaired so that the insulin prohormone itself (9 kDa) is secreted and can be found in blood, in addition to insulin and the C-peptide. What would a tryptically digested insulin peptide tell the researcher? Only a fraction of biology and pathophysiology would remain intact. In other cases, peptides may just differ by a few amino acids located at their N- or C-terminal end.

If one would now trypsinize these peptides, the biological peptide cocktail would be harmonized to an extent that makes a true peptidomics analysis literally impossible, especially if (semi-) quantitative information on peptide concentrations is also of interest. Consequently, no tryptic digestion (or by using any other protease) is performed during peptidomics workflows in order to conserve the precise molecular form of all peptides in the sample. This approach, called top-down-sequencing, is more challenging compared to analysis of short tryptic peptides. Only the endogenous analyte molecule, present in a complex peptide mixture, is analyzed and the resulting fragmentation pattern is searched versus databases. The analysis and matching of several tryptic fragments from one protein to the same database entry that would be regarded as confidence-building is therefore not possible. This has to be taken into account when publication of data is of interest since the concept to deny publication of “one-hit wonders” has become quite popular in specialist journals. In peptidomics, one will only find “one-hit wonders”!! In human blood, many thousand peptides are present in the mass range above 3 kDa, which is the upper limit for tryptic digests. For larger native peptides, several charge states such as 8^+ , 9^+ , 10^+ , or higher need to be selected and fragmented. The resulting daughter ion series at comparatively high-charge states need to be resolved to reveal their respective charge state. This is required in order to obtain a stable data template for sequence search by matching a characteristic MS pattern to a sequence database or *de novo* sequencing assigning the full amino acid chain step by step. This is usually done in MS/MS mode, by high-class MS/MS instrumentations such as Hybrid-ESI-Q-TOF mass spectrometers (QSTAR) or even Ion-Cyclotrons or Orbitraps. For smaller peptides up to ~ 4 kDa, the use of MALDI-TOF-TOF instruments is also feasible. In general, a combination of several setups will be necessary in order to increase coverage. A nice combination is the use of isobaric tags with such instruments, since quantification and identification are achieved simultaneously.

7.4 BIOINFORMATICS AND DATA MINING

A vast amount of data is generated during the many MS experiments in peptidomics. This can amount to tens and hundreds of gigabyte of raw data for a single study. In addition, other sources of data need to be integrated. Several distinct disciplines of bioinformatics are applied to solve the different challenges, which arise during such studies. Initially, appropriate MS-signal processing is required, ranging from baseline subtraction in individual spectra to mass calibration, quantitative (intensity-) normalization steps, and, in the usual case of LC-MS coupling, elution time normalization. At a certain stage, data reduction is necessary to be able to manage data by statistical tools.

Data curation is also needed to eliminate outliers and find clues and hints to improve weak parts of the wet lab process. The search for principal components has shown to be of great help, but is often discouraging. Whereas in other “omics”-type applications PCA seems well suitable to separate the groups of interest, this has not proven true in our hands. Frequently, the first principal components are not related to

the question “group A or group B” but to confounding experimental factors such as: repeated freeze-thawing, sample from center one or center two, first chromatography of the day, different operator (such as: Who took the blood sample?), and many more. Eventually, this information can be used to improve study design, as well as analytical procedures in the lab. Given the high susceptibility to changes, only a high degree of scrutiny will allow maturation to a level of quality where differentially measured peptides are derived from biology and not from one of the many artifact sources. Unfortunately, after 10 years of research we have not found a single suite of tools and recipes, or combinations there of, to fundamentally support our work, which is good enough to be trusted and to declare these problems solved. During this time we have evaluated many vendors and read so many wonderful and promising product descriptions. We recommend to meticulously document each step of the work and to be very cautious, even suspicious, in order to gain confidence in the data produced (and the conclusions drawn from that work).

As one of the first data analysis steps, data need to be analyzed in order to reveal their type of distribution, since this is necessary to check whether the statistical analysis that was planned can be performed by using such data. This is followed by more basic statistical work to interrogate means, standard deviations, p -values, and receiver–operator characteristics. Based on the common use of only a few samples with each sample giving rise to thousands of individual attributes, there is an inherent risk of overfitting models. Data mining is performed with a broad collection of tools, usually analyzing data in a supervised and unsupervised, high-throughput manner to discover information hidden within complex data. These data mining methodologies are often borrowed or adapted from microarray work. The integration of other data sources, such as from genomics, transcriptomics, or clinical data, is mandatory. The term “differential peptide display” may be misleading since the main analysis is not a subtractive analysis as indicated by the term “differential” but correlational analysis—either parametric or nonparametric.

Finally, the experienced bioinformatician will be of enormous help if this expertise is utilized in the experimental design of the studies. The design of a statistical and analytical analysis plan is considered one of the most important success factors for biomarker discovery. Experimental replication of the initial discovery study and the careful selection of controls—for instance, active (i.e., diseased, but not with the identical disease) and passive controls—is recommended.

7.5 DIFFERENTIAL PEPTIDE DISPLAY

The concentration of peptides and their relative relation in terms of half-life, distribution, production, and removal carries information about the morphology and physiology of an organism. Peptide composition of biological sources usually is highly dynamic. A comprehensive profiling of peptide patterns can uncover this information using MS as display tool [6]. A display system for peptides has to show that peptides are present in a sample. This will result in a qualitative analysis, and exclusive peptides that are present in only some samples can be found. Furthermore, since most

regulations in biology occur on the basis of concentration changes, a display system should also be capable to show the peptides in their respective concentration, resulting in quantitative analysis. In the case where no reference system with known concentration(s) is used, the quantification will be relative, resulting in fold changes instead of true concentrations. A good combination of qualitative and quantitative analysis may then allow selecting those peptides from the entire peptidome which are valuable for the respective analysis purpose. Originally, the method of analyzing and comparing peptidomes in a qualitative and quantitative way was called “differential peptide display.” This concept is per se not restricted to any specific combination of analytical instruments, but a best compromise of resolution, sensitivity, sample throughput, sequence identification, and quantification performance needs to be achieved. Apart from the ever-existing demand for highest sample quality, protein depletion and peptide extraction are the first crucial step in all workflows. As a second step, separation is performed by conventional HPLC. Then, analysis of peptide masses and quantity is done by MS, and data analysis is performed by appropriate bioinformatics tools. Finally, those peptides that are selected by biostatistics are sequenced by MS/MS to complete the workflow. Two alternative setups are recommended:

7.6 ID LC–MALDI

The MALDI–TOF–MS, nowadays also with instruments with TOF–TOF capability, has frequently been used for analytical profiling of peptides. Current improvements in mass accuracy, throughput due to increased laser frequency, and MS/MS on peptides from a sample spot allows for the fast, sensitive detection and identification of peptides in complex mixtures of hundreds of different peptides. Only nanograms up to a few micrograms of sample are required to create a detailed mass spectrum. Prior separation of the peptide extract into fractions, preferentially by RP–HPLC into 96 fractions, can increase the resolution and sensitivity to an extent that makes this combination attractive for the analysis of the plasma peptidome down to peptide hormone concentrations [6]. Several thousand native peptides can be measured in the mass range from 850 Da to 15 kDa with mass accuracies of better than 100 ppm from every single sample. Throughput allows for group sizes well into the range of 100 samples per group in a semiautomated workflow with an average coefficient of variation between 30% and 40%. A general characteristic of this workflow is the limited reproducibility of the process from spotting the sample to the target plate to MALDI measurement, because the coefficient of variation of ~20% for this part has a major impact on analytical quality. In addition, only a small percentage of the original sample is transferred into the mass spectrometer. If high sensitivity is required, 100–1000 μ L of plasma have to be processed due to the low transfer efficiency. In animal models, such as in mice, as well as in human clinical trials, plasma volume is always restricted. So this demand represents a serious drawback in many studies. The alternative use of microanalytical columns that directly deposit the effluent of the reversed-phase column onto a target plate has shown good individual data, but overall limited success, since column half-life is low, as is run-to-run and column-to-column reproducibility. We may

expect significant improvement for the LC–MALDI work once these key problems are solved.

7.7 2D CA–RP–LC–ESI–MS

This workflow follows conventional MudPIT-type experiments without the use of a protease to digest proteins. The peptide extracts are first separated into 5–25 fractions by strong cation exchanger, depending on the individual study. This type of separation is well-suited for complex biological samples, shows good reproducibility, and can be nicely interfaced to RP chromatography. The fractions from the first step are then analyzed by many conventional RP–LC–ESI–MS runs. Again, many thousands of peptides can be measured from a single sample. Depending on the MS instrument and the respective setup used, deconvolution of spectra can be problematic, since the m/z distribution for (larger) native peptides will result in several peaks per peptide, which appear at the same time and which are in overlay with other co-eluting peptides and their many different charge states. Resolution needs to be in the range of 10,000 HWMZ in order to allow the correct assignment of charge states. A higher degree of prefractionation will be beneficial in terms of resolution and sensitivity, but will lead to reduced reproducibility and multiple LC/MS runs per sample. Especially for this setup, the use of isobaric mass labels such as tandem mass tags (TMT) in combination with a high-class MS/MS instrument holds great promise.

7.8 APPLICATIONS

Much is known about the involvement of peptides in health and disease. Consequently, peptides are used as drugs (selection of peptide therapeutics in Table 7.1) and as diagnostics (selection of peptide diagnostics in Table 7.2). With the comprehensive peptide analysis in peptidomics studies, we can expect further progress in this field. In the search for relevant peptides in biological sources, much depends on preexisting know-how in order to design a successful study. In our experience, it is always very helpful to make a basis evaluation of the sample of interest in order to produce an

TABLE 7.1. Peptide Therapeutics

Compound	Trade name	Company	Indication	Sales (Mio US\$ 2002)
Insulin	Humulin	Eli Lilly	Diabetes	1004
Interferon- β	Betaseron	Schering	Multiple sclerosis	740
Parathyroid hormone	Forteo	Eli Lilly	Osteoporosis	>2000
Follitropin	Gonal-F	Serono	Fertility	410
Cyclosporine	Sandimmune	Novartis	Transplant rejection	1037
Natrecor	hBN	Scios	Heart failure	>50

Source: Modified from reference 8.

TABLE 7.2. Peptides as Diagnostics

Peptides	Disease
Osteocalcin, β -crosslaps	Osteoporosis
Insulin and C—peptide	Diabetes
Amyloid A β -42	Morbus Alzheimer's
LH/CG	Reproductive medicine.
Pr-GRP <small>small</small>	Lung cancer
β -2 Microglobulin	Renal failure, inflammation
ACT	Adrenal insufficiency
Angiotensin	Hypertension

Source: Modified from reference 8.

overview on that specific peptidome. Such work is usually done by LC–MALDI, and the researcher will quickly extract relevant information for the future work. Given the high diversity of samples to be used, specific peptide extraction procedures need to be used or developed. In most cases we could display several thousand native peptides that form the basis for discovery studies. As further technical feasibility, a subset of peptides was used to train the peptide identification performance and to obtain sequence information.

7.9 PEPTIDES AND PROTEASES

Peptidomics is closely linked to proteomics. The input of new peptides in a peptidome is usually generated by protease action on proteins from the proteome of the same sample. Due to their small size, peptides may also traverse from one compartment of an organism to another. This is especially the case in body fluids where individual organs secrete peptides into the fluid, sometimes in order to activate distant receptors. As a firm rule, peptides are produced from larger precursor proteins by proteases and peptidases that hydrolyze a peptide bond. The peptidome therefore represents a natural library of protease substrates and products. Human genomic sequence databases contain more than 500 proteases. For over 60 of these proteases, clinical development with protease inhibitors is ongoing. Proteases are one of the four large gene groups that are considered druggable. A large number of protease inhibitors are also present in the human genome, tailored to inhibit intrinsic as well as nonhuman proteases such as bacterial proteases. Proteases can be divided into five groups with regard to the respective catalytic center and show a wide range of specificities.

Proteases are the tools of biology which initiate, modify, and terminate intra- as well as extracellular functions by substrate cleavage, which can be highly specific and limited. This cleavage will, on the one hand, terminate protein function, but it will frequently deliver novel functional entities, which carry formerly hidden functional epitopes. It has been shown that functional epitopes are quite common in proteins and that they are used more than once, which means that almost identical stretches of amino acids occur in many proteins. In addition, several functional

epitopes can be co-localized within one precursor. They undergo specific activation at specific locations and in a timed fashion. The predominant mechanistic way of activation is proteolytic processing. In contrast to the activation status of intracellular proteins where phosphorylation and dephosphorylation by kinases and phosphatases comprise the dominant regulation principle, proteolysis is the key mechanism for extracellular regulation, where phosphorylation and dephosphorylation are not relevant.

The best-studied peptides are peptide hormones because they carry relevant biological activity. Their precursor molecules sometimes carry several distinct biological activities that are inactive after protein biosynthesis. A prime example is the pro-opio-melanocortin gene (POMC), which contains over 10 different peptide hormones and neuropeptides that are released by action of different prohormone convertases. The tissue-specific distribution of these peptidases is one key factor to determine which collection of hormones will be cleaved from the precursor molecule. This demonstrates that the release and the activation of individual peptides from precursors are triggered by specific proteolytic processing. Only at certain times and in specific compartments are they activated by different proteolytic processes. The regulated release of biologically active compounds from inactive (precursor) molecules is used to allow for the timely presence of a certain molecule in just the concentration that is required to fulfill a certain biological function. If that process fails, negative effects can happen. A failure in the activity of matrix metalloproteinases (MMPs) in the synovial layer of the knee joint will result in damaged cartilage, and production of wrong concentrations of angiotensin I will lead to hypertension. Other well-studied areas for the action of proteases and protease-cascades are coagulation (conversion of blood clotting factors from inactive to active states), digestion (the zymogens), blood pressure (angiotensin- and endothelin-converting enzymes), blood glucose regulation (dipeptidyl peptidase IV; processing as well as degradation of incretin hormones such as glucagon-like peptide 1 and gastric inhibitory peptide) and the above-mentioned MMPs, which are degradation enzymes responsible for the turnover of the protein matrix of connective tissue.

The complex regulation network of peptide and protein function by protease inhibitors contains many interacting partners, and most researchers have already had an impression by studying the blood clotting cascade. At least in the human system, this complexity seems to be rather the rule than the exception. Research with peptides always includes the awareness on proteases and protease inhibitors. Especially relevant for the clinical use of peptide biomarkers is the fact that proteolysis is irreversible, and that proteolysis frequently occurs in blood. It will be interesting to watch whether the action of proteases in plants will be of similar impact to understanding plant physiology as in mammals.

7.10 CONCLUSIONS

Peptidomics complements the analysis of gene expression products by including the group of native peptides to RNA, DNA, and proteins. Peptidomics can close

the gap between the analysis of metabolites, termed metabolomics, and the analysis of proteins, termed proteomics. It therefore forms an analysis platform that is capable to link static genotypes with dynamic phenotypes. The manifold functions of peptides in the organism, their role as products of degenerative processes, and their presence in cells, organs, and body fluids make peptides a promising target for analysis. At present, the peptidomes of many relevant biological systems have not been explored in-depth. In particular, the role of peptidomics in plants remains largely unknown. It will be a challenge to find out whether peptides, proteases, and protease inhibitors are of similar importance in plants as they are in humans.

7.11 FIVE-YEAR VIEWPOINT

The decoding of the human genome has enabled the world-wide proteomics community to progress with a speed that was unthinkable before. The combination of MS fragmentation of peptides with bioinformatics tools that search more and more completed genomic databases of many species allows the sequence analysis of hundreds, even thousands of peptides, within a single day. The application of improved analytical technology to the precious world of peptides will bring further benefit to the field. Higher sensitivity in combination with better quantification and structural identification of thousands of analytes will make peptidomics an integral part of systems biology approaches. A deeper understanding of peptidomics is expected, not only in humans but in all biological applications and species [7]. We expect that peptidomics will move from a niche application to a fully recognized “omics-type” technology. The challenge remains to use peptidomics technologies to understand functional variations in health and disease contexts. If this can be achieved, new opportunities for pharmaceutical but also for nutritional and nutraceutical applications will arise. We expect a future dominant role for peptides in body fluids and tissues as markers of disease [8].

Editors Note. Readers are referred to recent papers on the *M. truncatula* plant peptidome [9] and the *Arabidopsis* unannotated secreted peptide database [10] for further reading on the subject.

REFERENCES

1. Gaenshirt D, et al. 2005. *Peptidomics in Drug Development*, Edition Cantor Verlag, Germany.
2. Schulz-Knappe P, Schrader M, and Zucht HD. 2005. *Comb Chem High Throughput Screen* 8:697–704.
3. Schulz-Knappe P, et al. 2001. *Comb Chem High Throughput Screen* 4:207–217.
4. Petricoin EF, et al. 2002. *Lancet* 359:572–577.

5. Schulz-Knappe P, et al. 1997. *J Chromatogr A* **776**:125–132.
6. Tammen H, et al. 2005. *Comb Chem High Throughput Screen* **8**:735–741.
7. Crameri R, Schulz-Knappe P, and Zucht HD. 2005. *Comb Chem High Throughput Screen* **8**:807–810.
8. Schulte I, et al. 2005. *Expert Rev Mol Diagn* **5**:145–157.
9. Zhang K, et al. 2006. *J Proteome Res* **5**:3355–3367.
10. Lease KA and Walker JC. 2006. *Plant Physiol* **142**:831–838.

PART II

COMPUTATIONAL PROTEOMICS

BIOINFORMATICS IN GEL-BASED PROTEOMICS

Åsa M. Wheelock and Craig E. Wheelock

8.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

A third of a century ago, a novel method capable of separating 1100 proteins on one single PAG was introduced [1]. Although this powerful separation method could detect proteins with abundances as low as 1/10th to 1/100th percent of the total protein content in a cell, difficulties associated with visually estimating protein abundances limited its use to primarily qualitative applications. Since 1975, a steady stream of technical improvements has eliminated many of the original limitations [2] of 2-DGE to the point where semiquantitative studies have become possible. One of the major breakthroughs in this aspect was the development of powerful bioinformatics tools for image analysis and quantification. However, despite the fact that the resulting 2-DE analysis software were designed to reduce the experimental variance in order to enhance the biological variance of interest, the post-experimental software-assisted image analysis has been shown to introduce additional variance into the analysis [3]. In this chapter, we attempt to provide an overview of the main sources of post-electrophoretic variance in current 2-DGE analysis, as well as tools with which this type of variance can be quantified.

The first generation of computational approaches for analysis of 2D gels started to appear toward the end of the 1970s (Gellab, LIPS, Elsie, TYCHO) [4]. These early softwares were characterized by heavy user interaction and low automation, often requiring a great deal of programming and technical expertise by the user. A decade later, the introduction of graphical interfaces such as windows prompted a second generation of 2-DE analysis software (Elsie-4, Melanie, QUEST, Gellab-II). However, the original issues of non-user-friendly interfaces persisted, rendering this generation of 2-DE software inaccessible to non-computer scientists [4]. It was not until low-cost personal computers equipped with more user-friendly graphical interfaces and powerful processors became widely available in the late 1990s that the third, modern generation of 2-DE software was produced [4]. Several of these pioneers (Melanie II, CAROL, Z3, and MIR) [4] were subsequently developed into the 2-DE analysis software commercially available today, such as ImageMaster 2D Platinum™/Melanie™, PDQuest™, DeCyder™, Proteomweaver™, and Progenesis™. A major obstacle with this trend away from academic development toward heavy commercialization is that algorithms used in the products become proprietary information, resulting in a black box research approach. However, information regarding the workings of these softwares can often be inferred from the details of the algorithms in the original academic versions [5]. Efforts are still being made to produce custom systems [6], but since their use in the existing literature has been very limited, this line of products is beyond the scope of this chapter. Although the order of events may differ, the automated analyses of 2D gel images utilized in various commercial software programs generally include the following steps: (i) segmentation (spot detection), (ii) quantification, (iii) background adjustment, (iv) image warping, (v) registration (spot matching), and (vi) normalization. The challenges in computational 2-DGE analysis imparted by technical problems in the experimental technique such as artifacts or irregularly shaped or overlapping spots, as well as the purpose, strengths, and weaknesses of the algorithms used today, are highlighted below.

Spot Detection and Quantification

Quantification is often the primary goal in proteomics analyses and is, accordingly, a central point of 2-DE analysis software. The segmentation algorithms used in spot detection can be categorized into two classes, parametric and nonparametric. In parametric spot detection, the actual spot shapes in the 2D gel map are transformed into the “ideal” spot shape, generally by fitting to Gaussian parameters. The main advantage of this method is that the image can be replaced by a list of spot centers, which greatly reduces the complexity of the data file representing the image. For example, a 2D gel image containing 1000 spots is reduced to 28 kilobytes of data, equivalent to a file size of less than 1% of the corresponding pixel-based 2D gel map [7]. This technique greatly simplifies the subsequent spot matching step and was therefore used in earlier versions of 2-DE analysis software in order to overcome limitations in computational power. Parametric spot detection handles overlapping spots very well [8] and is still used in some products (e.g., PDQuest), although most modern programs utilize nonparametric, pixel-based segmentation algorithms where no constraints on the spot

shape are introduced [9]. Instead, quantification is performed through a summation of the pixel intensities localized within the defined spot area. The most widely used algorithm for defining the spot boundaries, termed “watershed,” has its origin in the geosciences and the behavior of water in mountain topography. The protein spots in a 2D gel represent the mountain peaks, and the areas where the water forms pools represent nonprotein spots, while the ridges—the watershed—are the protein spots [10].

Background Adjustment

The general idea behind background correction is to enhance the signal of the protein stain in the gel image through subtraction of background noise such as nonspecific staining or autofluorescence of the gel matrix. In order to distinguish the noise from the signal, spot detection is often performed prior to background subtraction, and non-spot areas are used for calculation of the background noise. Background subtraction can be performed locally or globally across the image. The simplest global strategy is “global constant” background subtraction, consisting of subtraction of all pixels below a set threshold of the maximum intensity. In contrast, global “morphological” algorithms take into consideration all the pixels in the gel that are not part of a detected spot when calculating the background (utilized in, e.g., PDQuest and Delta 2D) [3, 11]. In local background subtraction algorithms, the pixels in the immediate vicinity of the spot outline are utilized to calculate the background subtraction. Typically the average, lowest, or most frequent pixel intensity in the given region is defined as the background level and is consequently subtracted from each pixel value in the entire spot area (utilized in, e.g., Progenesis and ImageMaster). The 2D gel images are susceptible to the formation of speckles (salt-and-pepper noise), particularly using postelectrophoretic fluorescent stains. The speckles can be the result of staining of dust particles or SDS residues in or on the surface of the gel, and some software products offer additional filtering algorithms to reduce this type of noise [5, 7].

Image Warping and Matching

Registration of which protein spot corresponds to the same protein on a different gel image, generally referred to as spot matching, is another central point in 2D gel analysis. The original spot-based approaches described above, where a list of spot coordinates generated during spot detection was utilized in the alignment process, was complicated by the large gel-to-gel variations in spot migration patterns inherent in 2-DGE. The protein migration pattern can be distorted by a range of factors, such as (a) an inhomogeneous electric field during electrophoresis resulting from current leakage or (b) artifactual modifications of amino acids by constituents of the 2-DGE separation procedure. Accordingly, spot-based matching strategies (although still used in some programs such as PDQuest) have largely been abandoned and replaced by the use of raw image-based (also called pixel-based) registration strategies, where all the features in the gel image are utilized for matching [5]. Additional image warping

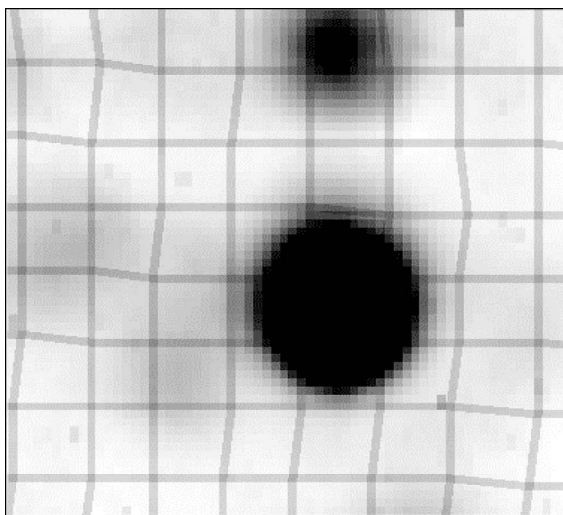


FIGURE 8.1. Display of the grid model used in many warping algorithms in order to counteract distortions in separation pattern that cause significant gel-to-gel variations even in replicate gels run under identical conditions.

algorithms, which deform the image in order to counteract geometrical gel-to-gel variations arising from experimental variation, have more or less become standard [11]. A common approach, illustrated in Figure 8.1, is the division of the gel images into a grid system where each cell is stretched to fit a master gel [12]. The warping results in both faster and improved accuracy in spot matching. In the past, the warping step has generally been incorporated as an automated algorithm performed in association with the matching step. Recently, a new workflow, perhaps representing a fourth generation of 2-DE analysis software, gives the warping algorithm a more central position in the analysis. In these products (e.g., Delta 2D and SameSpots), warping is performed prior to any image analysis and is said to drastically reduce user time as well as offer improvements in accuracy and reduced subjectivity of the subsequent matching.

Normalization

Despite extensive technical improvements of the 2-DGE technology over the years, considerable experimental variance still prevails. A large portion of the variance can be ascribed to inefficiencies in protein transfer both during rehydration of the first dimension and from the first to the second dimension [13]. Variations in protein staining efficiency, both when using covalent tags or after electrophoretic staining, are other major sources of experimental variance. To counteract the confounding effects upon the ability to quantitatively determine the true biological variance, various techniques for normalization have been developed and subsequently incorporated as standard features in 2-DE analysis software. One of the first normalization methods

utilized in 2-DGE was “total spot volume (TSV) normalization” [14]. In this “global ratiometric” normalization technique, the absolute spot volume is converted into a relative quantity through dividing the individual spot volume of each spot of interest with the sum of all the spot volumes detected in the gel image. A more common variant is “total valid spot volume (VSV) normalization,” where global ratiometric normalization is performed using the sum of a select set of spot volumes (e.g., those validated across all replicate gels). VSV is slightly more robust, because it decreases the influence of missing values. Both of these methods are provided in commercially available 2-DE analysis software.

In recent years, the development of multiplexing capacity (see DIGE in Chapter 2) has resulted in more sophisticated normalization methods involving “direct ratiometric” normalization. The availability of pI-matched, spectrally separated fluors (Cy₂, Cy₃, and Cy₅) allows multiple samples to be co-separated in essentially identical patterns on the same gel, and normalization can be performed through dividing each individual spot volume of one fluor with the corresponding spot volume of a different fluor on the same gel. Most efficient normalization is achieved when one of the CyDyes is used to label an internal standard, ideally created through pooling of all the samples to be analyzed in the study. This assures that all protein spots are represented in the internal standard, thus facilitating direct ratiometric normalization of all protein spots [15]. In addition, the utilization of an internal protein standard allows correction for technical gel-to-gel variability between gels, which significantly improves the statistical aspects of the quantitative analysis. However, the use of a pooled internal standard requires the sacrifice of up to one-third of the sample. Additional requirements for a pooled standard include that all samples to be included in a given analysis are collected prior to initiating the experiment, something that may prove difficult to fulfill in clinical studies where samples may be collected over an extended period of time. In such circumstances, an alternative internal standard approach similar to that utilized in ALIS (Alexa-fluor-labeled internal protein standard) methods may be advantageous [16]. The ALIS method is a cost-effective alternative to DIGE where a tissue sample resembling the sample of interest—rather than a pooled standard—is fluorescently labeled and utilized as the internal protein standard. Total protein is visualized using a postelectrophoretic fluorescent stain spectrally separated from the ALIS (e.g., SYPRO Ruby), thus facilitating multiplexing. However, because the ALIS approach does not result in superimposable separation patterns of the internal standard and the sample proteins, a global ratiometric normalization method is used. The main weakness of global ratiometric normalization methods is the lack of ability to correct for local differences in background and separation patterns. Direct ratiometric normalization is dependent on superimposable separation patterns and thus sensitive to missing values. The 2-DE analysis software tends to be better equipped to handle direct ratiometric normalization for multiplexing. Software marketed for multiplexed analysis is also equipped with algorithms capable of correcting for the dual spot migration pattern that is a result of minimal labeling strategies. The use of minimal labeling reduces problems associated with low protein solubility imposed by covalent labeling with the hydrophobic fluorophores used in multiplexing, and it implies that only a few percent of the total amount of proteins are

labeled with CyDye. However, one should keep in mind that even if the algorithms can estimate the location of the “invisible” bulk of the protein during spot picking, problems due to overlapping spot patterns still may arise during subsequent protein identification through MS.

8.2 METHODOLOGY AND STRATEGY

In recent years, the notion that 2-DE analysis software may introduce additional variance into the analysis has been brought to the light [17]. Unfortunately, the manufacturers of these products generally do not provide any performance evaluations, and the research community is thus forced to rely on the few evaluations performed by other users [for review see reference 3]. Furthermore, no consensus concerning standardized tests for evaluating the overall performance has been reached in the proteomics community, which greatly complicates the issue of comparing different studies and products. The quality of the gel images as well as the particulars of the experimental protocol used, such as the choice of protein visualization method, may have a drastic effect on the performance of the software. Nevertheless, in order to determine the power of study results, it is necessary to evaluate the performance of spot detection, matching, and quantification. This chapter suggests methods for evaluating 2-DGE analysis data in terms of both the quality of automatic spot matching and how to quantify the variance induced by specific 2-DE analysis software. The basic protocol for how to perform software-assisted 2D gel image analysis is covered elsewhere [18]. For the purpose of this chapter, a standard set of images will be employed. These images are freely available for download from www.pulmonomics.net, enabling readers to regenerate the described analysis as well as employ the same images in evaluating other software products not covered in this review.

Image Acquisition

Image acquisition is often trivialized in 2-DGE analysis, even though reproducible image acquisition is as essential as any other step in minimizing variance. Image acquisition instrumentation can roughly be divided into two categories: (i) via photography, where excitation occurs through illumination with a constant light source (UV or Xenon lamp) followed by detection using a cooled CCD camera, and (ii) through scanning, where a laser is utilized for excitation at specific wavelengths (e.g., Ar: $\lambda_{\text{ex}} = 488$ nm; ND-YAG: $\lambda_{\text{ex}} = 532$ nm; He-Ne: $\lambda_{\text{ex}} = 633$ nm) and a photomultiplier tube (PMT) is utilized for detection [for more technical details see reference 19]. In both types of instrumentation, the use of emission filters is central to reduce background and increase specificity. In the case of CCD-based systems, emission filters may also be used to simulate the specific excitation wavelengths of the laser. Emission filters can either be of long-pass (LP) or band-pass (BP) type (Figure 8.2). Long-pass filters exclude all light of wavelengths below the given limit, and they are generally used to exclude the light source itself. For example, an LP540 filter used in conjunction with an ND-YAG laser prevents light from

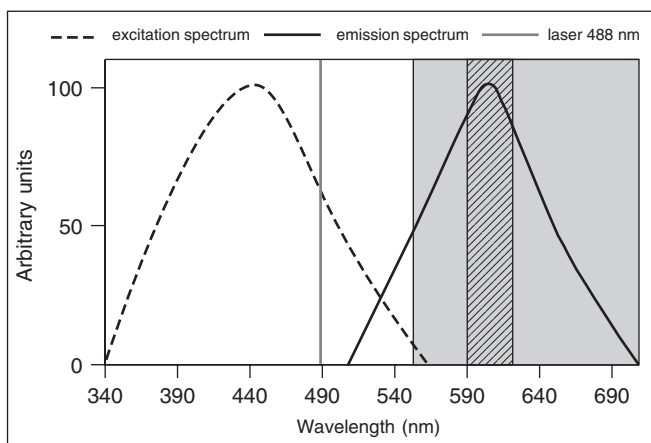


FIGURE 8.2. Illustration of the use of long-pass (LP) and band-pass (BP) emission filter, in conjunction with SYPRO Ruby protein staining. The dotted line represents the excitation spectra, while the solid line represents the emission spectra. The use of a LP filter (here 555 nm) allows cumulative collection of all emission above a given wavelength (entire gray area), resulting in high sensitivity. The use of a BP filter (here 610 BP30) selectively collects the emission from a narrow range, here 30 nm centered around 610 nm (striped area), resulting in high specificity.

the laser itself from confounding detection of the emission spectra. The LP filter thus gives a high yield, because the entire emission curve of the fluorophore is collected. However, it also detects the cumulative nonspecific emission of the range; hence it does not necessarily result in a higher signal/noise ratio, because the specificity of the signal is reduced. Most biological materials autofluoresce around 500–600 nm, and accordingly they cause more problems with Ar or NAD-YAG lasers (Cy3, SYPRO Ruby, Deep Purple) than with He–Ne lasers (Cy5, Alexa₆₃₃, Alexa₆₄₇). The nonspecific emission can be reduced through the use of a band-pass filter. For example, a 560 BP30 filter used in conjunction with a ND-YAG laser limits the detection to a narrow window of the emission spectrum centered on/around 560 nm (Figure 8.2). BP filters are a necessity when utilizing multiplexing because multiplexing reduces “bleeding” between different flours used on the same gel, given that the flours used for multiplexing have sufficient spectral separation to avoid quenching.

Laser scanners have traditionally been ascribed a higher resolution and a broader dynamic range than CCD-based systems. However, the 16-bit format utilized in most modern CCD cameras provides, at least theoretically, a dynamic range of over four orders of magnitude, which is typical for most laser scanners. In addition, newer systems (such as Perkin Elmer’s ProXPRESS) exhibit 33- μ m resolution [20], which approaches that of laser scanners. The Xe/UV light source used in this system also makes it compatible with most fluorescent protein stains on the market [20]. The choice between the two types of systems may not have as big an impact on the

quality of the image acquisition as it used to, although the higher excitation energy of the laser will likely result in an improved limit of detection.

Regardless of the instrumentation used, the image acquisition step has the potential of introducing variance into the analysis. Unfortunately, no rigorous studies evaluating the actual variance have yet been reported. Ideally, this should be tested for each system through repeated scanning of a single gel and through quantification of the signal with robust imaging software (e.g., ImageQuant; Nonlinear Dynamics, Sunnyvale, CA). Furthermore, the PMT setting is bound to have a nonlinear quantitative effect on spot volume and total amount of protein spots detected. This has been investigated in detail for the Typhoon 9400 laser scanner (GE Healthcare, Uppsala, Sweden), showing an exponential relationship between PMT setting and pixel value for all fluors tested (Cy₂, Cy₃, Cy₄, SYPRO Ruby) in addition to a constant, gel-specific difference [21]. Fluorescent residues from previous scans that adhere to the glass platen, particularly when using postelectrophoretic stains, can also contribute substantially to the variance in background fluorescence. This source of error can be avoided through careful cleaning of the glass platen with an appropriate solvent (e.g., propanol). Remember to check that the solvent is compatible with the material of the glass platen prior to use.

Evaluation of Automatic Spot Detection Algorithms

All 2-DE analysis softwares are equipped with some form of automatic spot detection and matching tool. Unfortunately, frequent mistakes in both spot detection and matching are made due to discrepancies in the spot migration pattern, staining artifacts, and so on. These mistakes are generally corrected through subsequent manual revision by the user. In addition to being extremely time-consuming, this process also introduces a large degree of subjectivity into the results, because the user has to make active choices both in terms of whether a detected spot truly is a spot and in terms of evaluating correct matching of spots across replicate gel images. For these reasons, it is desirable to optimize the automatic analysis step with the various user-defined settings available in the software. The results can then be evaluated using free-response operator characteristics (FROC) curves. The FROC (or receiver operating characteristic, ROC, depending on application) curve, a commonly used concept in evaluating the performance of medical diagnosis or pharmaceuticals [22], was introduced as a tool for evaluating 2-DE analysis software by Rogers et al. [8]. The FROC curve represents the relationship between sensitivity and selectivity of the test. In gel-based proteomics, sensitivity is the capability of detecting true spots, while selectivity is the capability of excluding artifacts from being detected as spots. In the standard format, a FROC curve is plotted as the TP fraction: $TPF [(\# \text{ of correctly detected spots})/(\text{total } \# \text{ of spots in the image})]$ versus FPF [the FP fraction ($\# \text{ of artifacts detected as spots})/(\text{total } \# \text{ of artifacts})$]. However, because it is difficult to calculate the total number of artifacts in a gel image that potentially may be detected as a spot by the algorithms, it is more practical to either plot the actual number of detected artifacts (FP) in the FROC graph or use the highest number of detected artifacts as “total.” In the examples provided in the results section, we have chosen the latter.

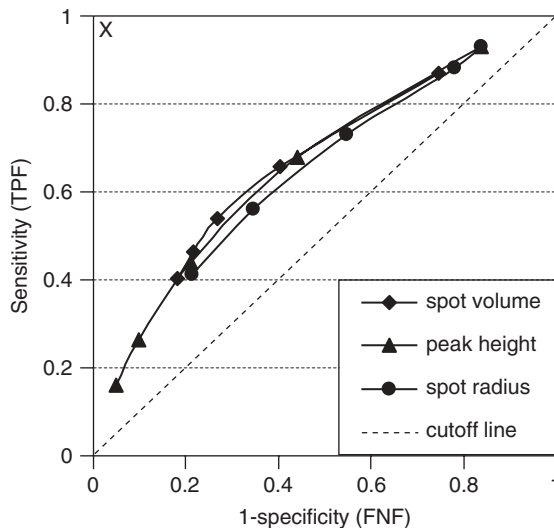


FIGURE 8.3. FROC curves were generated in order to evaluate the various filtering options available in Progenesis 240 following analysis with the SameSpots software. A range of values for minimum spot volume, peak height, and spot radius were used. The diagonal line represents no discrimination, and the "X" mark in the upper left corner represents the ideal result achieved through manual spot editing. None of the filters were particularly effective in filtering out artifacts (FP) from true spots (TP), although spot volume appeared to offer slightly better selection criteria.

Regardless of which method you choose, your ideal value will be located in the upper left corner of your graph (Figure 8.3).

Evaluation of Software-Induced Variance

The method presented here is designed to quantify the amount of variance introduced by a 2-DE analysis product and is an adaptation of a previous study by Wheelock and Buckpitt [17]. All images discussed are available for download from www.pulmonomics.net. Alternatively, new images can be created as described below. Pick one representative 2D gel image from an analysis set and copy it five times under different file names. Crop the images one by one using appropriate imaging software (e.g., ImageQuant). It is crucial that the cropping is performed manually with the mouse, because any tool designed for copying the cropping area between images will result in the exact same cropping of the copies. Instead, a gel set with a slight difference in image boundaries should be produced, because this slight shift in boundaries generally results in the software failing to recognize that the replicates in fact are the exact same image. The point of this procedure is to produce "replicates" where both experimental and biological variance has been excluded, in order to reveal any variance caused by the software analysis itself. Next, perform a

quantitative analysis of the “identical replicate” gel set. Make sure to test all available background adjustment options, because these may have a significant impact on the quantitative results [17]. Review the spot detection and matching manually to assure its accuracy, and then export a sufficient set of spot quantities to Excel. Calculate the average spot volume and the coefficient of variance (CV) for each of the matched spots across the five replicates and plot them in a bar graph. By ordering the values according to average spot volume, you can discover trends in the variance related to spot size. Repeat the analysis with results from different software programs, normalization methods, or background subtraction methods. Please keep in mind that while the exclusion of background subtraction can reveal if a certain algorithm introduces bias in the analysis, the omission of background subtraction will result in lower CVs as the total spot volume remains larger for all spots.

The strategy described above relies on a certain amount of sensitivity to the location of the image boundary in the 2-DE analysis software. Some programs—for example, ImageMaster 2D Platinum [3]—may be more robust in this matter, which makes it harder to “fool” the software into treating the individual copies of your gel image as replicates. In this case, you can repeat the analysis of a small area of your gel image and can match a subset of spots manually in order to perform your variance analysis.

Analysis of Data Distribution and Variance

Most statistical methods used to determine significant alterations are based on the assumption that the data are normally distributed, yet distribution analysis of 2-DGE data is often neglected. Two types of distribution analyses ought to be performed on each data set: (i) the distribution of spot volumes of each individual protein spot across replicate gels (“spot volume distribution”) and (ii) the distribution of the resulting variances of the spot volumes across the replicate gels (“variance distribution”)—that is, the values resulting from the “evaluation of software-induced variance” described above (Figure 8.4). The distribution pattern can be visually assessed through a histogram (Figure 8.4, upper panel) or a Q–Q–normal plot, in which the sample quantiles are plotted against the theoretical quantiles in the corresponding normal distribution (Figure 8.4, lower panel). A linear correlation implies that the sample is normally distributed, and a formal goodness-of-fit test should be performed for verification. The Shapiro–Wilk test was developed for small sample sizes [23], and it is a suitable choice for omics experiments (i.e., large-scale data approaches where the number of replicates typically is low [16, 17, 24]). Based on the null hypothesis that the data are normally distributed, the test calculates the correlation of the points in the Q–Q–normal plot. A rejection of the null hypothesis ($p < 0.05$) thus implies a non-normal distribution, and appropriate transformation of the data should be considered. The most commonly used transformation for 2-DGE data is log-transform. However, Kreil et al. [25] have reported that log-transform may lead to inflated variance at low signal levels [25]. The many similarities between mRNA and protein global expression analyses have prompted the exploration of applying transformations common in microarray experiments on 2-DGE data, and the successful use of Arsinh transformation to achieve

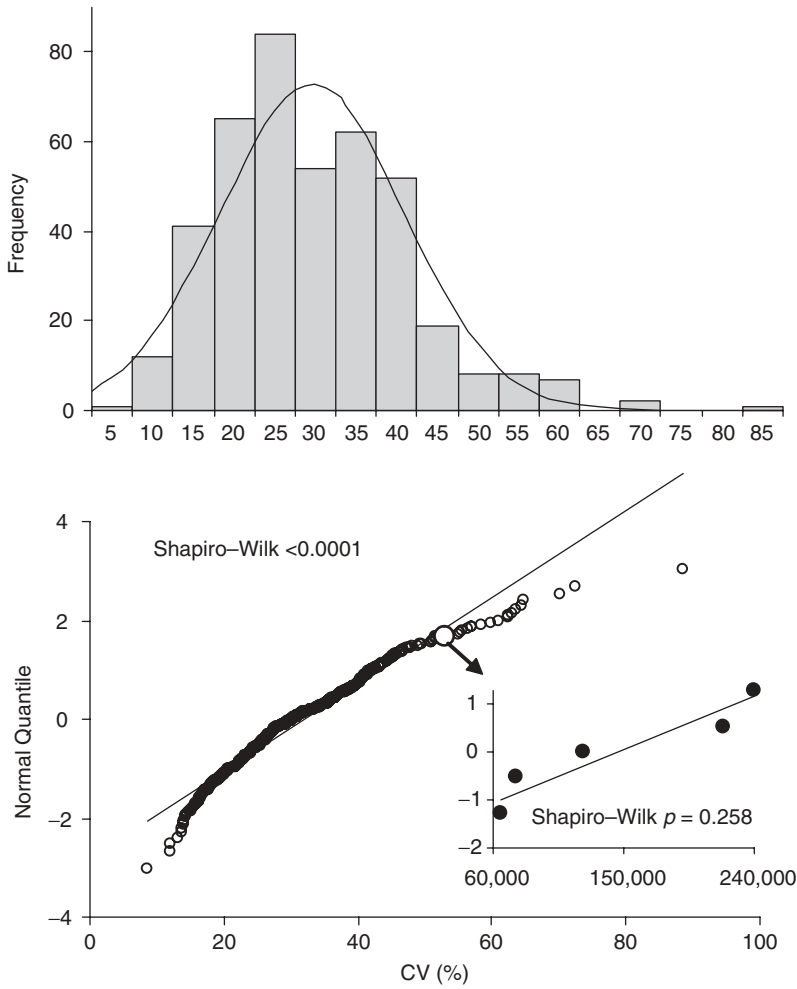


FIGURE 8.4. The bar graph in the top panel shows the distribution of the variance in spot volumes resulting from an analysis of five replicate 2D gel images (i.e., including technical variance). The overlaid curve in black shows the corresponding normal distribution. The main graph in the bottom panel shows the Q–Q–normal plot for the same data set. Formal goodness-of-fit test using Shapiro–Wilk W test revealed that the data set has a significantly non-normal distribution. The inserted graph shows the Q–Q–normal plot for the distribution of spot volumes for one of the spots used in the variance distribution plot analysis. The Shapiro–Wilk W test confirms that the spot volume data are normally distributed.

a normal distribution has been reported by two different groups [24, 26]. “Quantile normalization” has also been reported as a successful scaling strategy, particularly for SYPRO Ruby-stained gels [14].

Statistical Analysis

A typical 2-DGE experiment has an identical advantage and disadvantage: A significant amount of data can be generated from a single experiment. Following the collection of these large datasets, it can be challenging to extract the meaningful biological data in a statistically appropriate context. A typical problem is not having sufficient degrees of freedom in the experimental design—in other words, a high number of variables (protein spots) and a small number of measurements (replicate gels) [27, 28]. Standard univariate statistical approaches (e.g., Students-, *t*-test) could be appropriate if the experiment was designed to only examine alterations in a few proteins. However, univariate tests are inherently sensitive to type I error (FP); and because the investigator usually wants to analyze as many proteins as possible simultaneously, a multivariate or inductive approach for pattern recognition is often more appropriate. In this case, the experiment is designed as a hypothesis generating approach instead of hypothesis-driven. Pattern recognition can be divided into supervised and unsupervised approaches. Unsupervised pattern recognition is useful to determine if data fall into distinct groups where the aim is to detect data similarities, and subsequently no particular biases exist in terms of the group identifications. Examples include cluster analysis consisting of similarity measurement (e.g., correlation coefficients, Euclidean distance, and Manhattan distance) linkage (e.g., nearest neighbor and furthest neighbor) and hierarchical clustering (e.g., dendrograms or tree diagrams) [27, 29]. Supervised pattern recognition on the other hand attempts to answer a precise question as to the class of an unknown sample and therefore requires a training set of known groupings to construct the model.

A 2-DGE experiment usually requires an unsupervised approach, of which PCA is one of the most common methods. The use of PCA is advantageous, because the techniques are designed to transform a large number of possibly correlated variables into a smaller number of uncorrelated variables, or PC. These analyses can be thought of as essentially variable reduction and are used to identify hidden structures in a dataset. Because many 2-DGE experiments consist of literally thousands of individual proteins, which are often compared over multiple dosing regimens and time-frames, a method capable of dramatically reducing the number of variables to a more manageable data set is advantageous. PCA reduces the dimensionality of the data set through a series of transformations that result in a low-dimensional plot of the data [29]. In the analysis, data are structured such that the rows are the samples (in 2-DGE: the gels) and the columns are the variables (in 2-DGE: the protein spots). Many 2-DE analysis software packages include multivariate statistical packages, making the analyses straightforward. However, these statistical packages are often rather simplified, and since the spot volume data can be obtained as standardized output from most

2-DE software programs into Excel or text format, it may be advantageous to use a specialized statistical package for the analysis.

Output from the PCA analysis includes a series of scores and loadings, in which the relationship of the PCs to the samples is described by the scores and that to the variables is described by the loadings. A PC is a linear function of the original variable that can be thought of as a vector in multidimensional space, with each variable representing an axis [29]. The number of significant PCs is ideally equal to the number of significant components. The first PC describes the majority of the variance, the second PC describes the next greatest portion, and so on. The scores have as many rows as the original data matrix and the loadings have as many columns as the original data matrix. The size of each PC is given by the eigenvalue, which can be defined as the sum of squares of the scores. The sum of all nonzero eigenvalues for a data matrix equals the sum of squares of the entire data matrix. The resulting data are often plotted to examine the biological meaning. One of the simplest plots is that of the score of one PC against another. A scores plot can display clustering of distinct groups within the dataset. A loadings plot can then be used to display which loadings (i.e., protein spots) are driving the observed clustering in the scores plot. A biplot superimposes a scores plot and a loadings plot onto a single graph.

All chemometric methods are influenced by the method employed for data preprocessing. An understanding of data preprocessing is essential for correct interpretation of the output from statistical packages. The simplest transformation is of course none at all, and the raw data can be employed in the analysis. However, mean-centering, in which the mean of each variable is subtracted from each variable, is very common. This transformation shifts the scores plot such that it is centered at the origin. However, it can also affect the relative positions of the variables in both the scores and loadings plots. Mean-centering can often reduce the size of the first eigenvalue and influence the apparent number of significant components in a dataset. Another common method for data scaling is standardization, which is performed following mean-centering. Standardizing the data involves dividing each variable by its standard deviation, changing the covariance matrix to the correlation [29]. This transformation places all of the variables on approximately the same scale, enabling low values to assume equal significance as high values. If standardization is not performed, then the PCA will be dominated by the most intense (highest abundance) components. There is no general guidance as to whether to use centered or raw data when determining the number of significant components, the most appropriate method being dependent on the nature of the experiment. However, the biological significance of the transformation must be considered when analyzing the data. If the data are standardized, that means that changes in small abundant proteins will contribute significantly to the data analysis. Because these proteins are often at the detection limit, they may consist of some artifacts and should be manually confirmed. However, it is often these low-abundance proteins that are of interest, explaining why the majority of studies standardize the data.

8.3 EXPERIMENTAL RESULTS AND APPLICATIONS

In order to demonstrate how the methods described in the methodology section above are applied to an authentic set of 2D gels, we have performed an extension of the analyses carried out in a previous study comparing the performance of two 2-DE analysis software, PDQuest and Phoretix 2D Expression (now PG200) [17]. In brief, the data set consisted of five technical replicates of an airway epithelial sample separated on pI range 4–7 and a Duracryl gel matrix, and it was visualized with SRPGS as previously described [17]. In addition, an “identical replicate” set was generated as described above through copying of one gel image and subsequent individual cropping of the gel using ImageQuant. These images were selected for the study based on their high background staining and substantial speckling, because these characteristics challenge most 2-DE analysis programs. In the examples below, we analyzed these gel images with the SameSpots/PG240 software (Nonlinear Dynamics, Newcastle, UK).

Evaluation of Automatic Spot Detection Algorithms

Following user-guided warping and alignment of the five replicates, automated spot detection was performed through the SameSpots algorithm. A total of 1143 spots were detected and correctly matched across all gels. Manual review of the spots was performed to distinguish the TPs (520 spots) from the FPs (623 spots). All 1143 spots were exported to PG240, and the efficiency of the spot filtering tool in discriminating true spots (TP) from artifacts (FP) was evaluated. Following spot filtering, the TPF and FPF ($= 1 - \text{TNF}$) were calculated as described in the methodology section and were plotted in a FROC curve. The results from three different spot filtering criteria at a range of different settings are shown in Figure 8.3 (spot volume $> 50,000$ – $250,000$, peak height > 500 – 2500 , and spot radius > 4 – 12). None of these spot filters resulted in good discrimination of true spots from artifacts, as evidenced by the proximity of the graphs and the cutoff line.

Evaluation of Software-Induced Variance

In contrast to the previously performed study using PG200 and PDQuest [17], no manual editing using semiautomatic editing tools was required following spot detection and matching with SameSpots, because all spots were correctly matched. Quantitative evaluation of the software-induced variance as well as the experimental variance was evaluated on 416 manually reviewed true spots using all five background subtraction algorithms (no-background-subtraction, lowest-on-boundary, average-on-boundary, mode-of-non-spot, and progenesis). The resulting CVs for the five identical replicates using progenesis background subtraction are shown as a bar graph in Figure 8.5 (upper panel). The results for all algorithms are shown as whisker box plots in the lower panel (Figure 8.5). The complete omission of background subtraction results in the lowest variance. However, inclusion of the background noise in the average spot volume used to calculate the CV results in an overall lower CV than the corresponding variance would cause after background subtraction. Accordingly, a transformation of

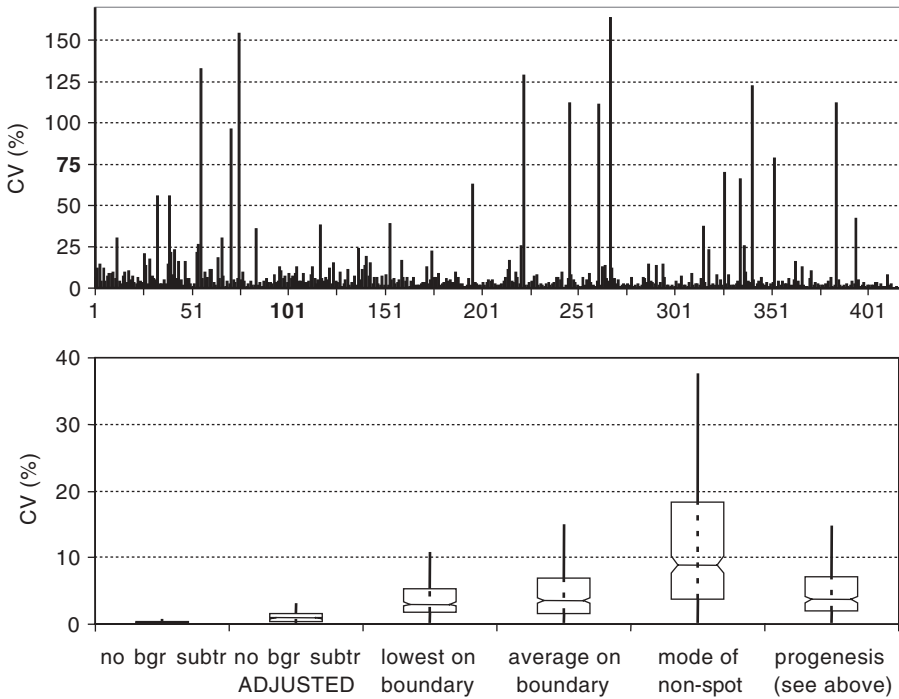


FIGURE 8.5. The top panel shows the variance in spot volumes across five copies of the same gel image following analysis with SameSpots, using the Progenesis background subtraction algorithm. Each bar represents the CV of one protein spot across the five identical replicates, and the bars are ordered according to increasing average spot volume. The bottom panel shows a box plot representing the CVs for the same experiment for each of the available background subtraction algorithms (lowest-on-boundary, average-on-boundary, mode-of-non-spot, and progenesis). In addition, the results from complete exclusion of background subtraction (no bgr subtr), as well as an adjustment of these results to correct for the decrease in CV caused by the larger spot volumes resulting from exclusion of background subtraction, are shown. Each box contains the interquartile range with the mean marked inside, and the whiskers show the range of CVs following exclusion of outliers.

the no-background-subtraction results was performed through dividing the spot volumes with the ratio of the no-background-subtraction and lowest-on-boundary spot volumes (Figure 8.5). The trend remains the same also after transformation, emphasizing the importance of evaluating the choice of background subtraction method because the quantity and nature of the variance in a system has direct implications on the statistical power. Considering that the biological variance within the control group is 40–50% [30], an additional technical variance of 10–20% caused by the analysis software will have a significant effect on the number of replicates needed. Based on a statistical confidence of 80% power and a 0.05 p -value, an overall variance of 50% would require $n = 3$ –4, while a variance of 70% would require $n = 8$ [30].

The same analysis was also performed for the authentic 2D gel replicate set (i.e., including technical variance). The software-associated variance accounted for 12–16% of the total variance for all background subtraction methods offered in the SameSpots software—except mode-of-non-spot, which caused for 35% of the total variance. These results correspond well to the previous studies performed with Expression (PG200) [17]. In similar studies performed on DIGE gel images, the DeCyder software appeared to introduce 30% of the total variance (reported as the unexplained variance) [31], which corresponds well both with previous results from PDQuest [17] as well as the results presented here using the mode-of-non-spot algorithm.

Analysis of Data Distribution and Variance

While the reason for the spot volume distribution analysis generally is intuitive, the purpose of the latter may not be as obvious. In quantitative omics analyses, we strive to determine significant biological alterations in expression levels. Due to the sheer number of protein expression levels to be evaluated, it is not always practical, or even appropriate, to perform rigorous statistical testing on each protein spot. Instead we utilize a cutoff level to select proteins of interest. By convention, 1.5- fold or 2-fold changes in expression levels have been utilized as cutoff levels. However, the assumption underlying the concept of the cutoff level strategy is that all of the spots in the data set have a similar behavior in terms of variance and distribution. To determine if a given cutoff level is appropriate, we thus have to determine that the variance for individual spots come from the same population—that is, have a normal distribution. The results from the previous section were analyzed according to the methods described in Section 8.2. It is of great importance to include “all” spots in an evaluation of the overall variance of a system and not just a few “well-behaved” spots. The results from the authentic replicate set (i.e., including technical variance) using lowest-on-boundary background subtraction are displayed in Figure 8.4. By visual inspection of the bar graph (upper panel), the distribution appears normal. However, formal goodness-of-fit testing revealed that the data set has a significantly non-normal distribution (lower panel). The inserted graph shows the Q–Q plot for the distribution of spot volumes for one of the spots used in the variance distribution analysis. The Shapiro–Wilk test confirms that the spot volume data are normally distributed.

8.4 CONCLUSIONS

Historically, the variance induced by computational analysis in gel-based proteomics has been considered neglectable in comparison to the experimental variance. However, the immense technical improvements in recent years have decreased the experimental variance to a point where variance arising from postelectrophoretic analysis becomes prominent, and the portion attributed to the software alone may exceed 30% of the total technical variance [17]. Accordingly, the choice of software product or algorithms offered by the chosen software may have a profound effect on the outcome of the study as well as on the time spent on computer analysis [6, 17]. The lack of a

standardized test for the evaluation of software performance makes it difficult to objectively compare the quality of different programs. Toward this end, we suggest that sets of standardized tests in conjunction with standardized sets of gel images be made available to the research community to serve as a benchmark for software and algorithm comparison. The images discussed in the current work are available for download from www.pulmonomics.net. It is our intent that this work spurs the research community and commercial interests to expand on the points raised in this chapter. It is important that quantitative comparisons of individual software packages are performed in order to evaluate program efficacy as well as enable researchers to quantify and interpret increasingly small variances in biological data sets. We have provided a set of tools for determining the quality of a 2D gel image analysis, both in terms of reproducibility and in terms of quality of spot detection and matching. These tools may also be utilized to compare the performance of different products or to optimize the user-defined settings within a program.

8.5 FIVE-YEAR VIEWPOINT

In spite of the revolution that has occurred in the quality of 2-DGE-based separation techniques since its introduction 32 years ago, gel-to-gel variability persists in 2D gel analysis. As such, fast and robust image processing is a crucial step in quantitative gel-based proteomics, and the fast development in 2-DE analysis software in recent years has removed some of the major stumbling blocks in the field. Future efforts are likely to focus on the improvement and evaluation of previously neglected areas such as background subtraction algorithms and statistical analysis as well as decreasing overall analysis time. Furthermore, standardized sets of 2D gels as well as performance benchmarks for the evaluation of 2-DE analysis software need to be established. Attempts have been made to produce sophisticated sets of synthetic gel images that reflect the true characteristics of authentic 2D gels [8, 9]. Rogers et al. [9] created a novel model for the creation of synthetic protein spots based on a training set of authentic 2-DGE spots [9]. Unfortunately, no background was introduced to these images; and as evidenced by the results in Figure 8.5, the background subtraction algorithm can be one of the main sources of variance introduced by the 2-DE analysis software [8]. Until an ideal set of synthetic gel images reflecting all aspects of the authentic 2D gel has been constructed, diverse sets of authentic 2D gels representative of common distortions (high background, speckles or irregularly shaped spots [17]) ought to be utilized. As user awareness of the effects of varying algorithms upon software performance grows, manufacturers may finally be forced to provide detailed information on the performance of their products. These types of data would greatly assist in software acquisition, enabling potential buyers to evaluate which software is most appropriate for their intended applications. More flexibility and user-defined functions will be on demand as alternative normalization methods are developed. Recent reports that the state-of-the-art fluorescent protein stain, SYPRO Ruby, may have polynomial rather than linear correlations to protein quantity may increase the demand for user-defined quantitation curves in commercial 2-DE software [16]. Furthermore, the limitations of algorithms for background subtraction may be resolved

through the introduction of time-resolved fluorescence. The delayed measurement of a fluor's emission excludes autofluorescence from plates, reagents, or cell debris, which generally have very short life spans (low nanosecond range) and thus eliminate a major source of background noise. In contrast, fluorescent protein stains have very long life spans (high nanosecond to microsecond); and delayed, cumulative detection over time may improve sensitivity [32, 33]. The use of both covalently labeled fluors (CyDyes and Alexa-dyes) [33] and ruthenium chelates [32, 33] have been utilized in time-resolved fluorescence applications in related fields. However, the lack of this feature in modern 2D gel image acquisition equipment is currently prohibiting the use and development of time-resolved fluorescence in 2-DGE.

The growth in data acquisition combined with increased quantification capability will continue to expand. These large bodies of data will enable researchers to further understand complex cellular processes and move the field another step closer to comprehension at the organism level. Proteomics data will enable researchers to model and understand interactions in a cell and to predict the effects of fluctuations upon other intracellular processes. These effects will be seen through, for example, the application of proteomics to web-based models of cellular metabolism such as an electronic cell [34]. Proteomics research will assist in the constant march toward true systems biology that will revolutionize personal medicine and fundamentally shift our understanding of biological processes.

REFERENCES

1. O'Farrell PH. 1975. *J Biol Chem* **250**:4007–4021.
2. Gorg A, Weiss W, and Dunn MJ. 2004. *Proteomics* **4**:3665–3685.
3. Wheelock AM and Goto S. 2006. *Expert Rev Proteomics* **3**:129–142.
4. Appel RD, et al. 1997. *Electrophoresis* **18**:2724–2734.
5. Dowsey AW, Dunn MJ and Yang GZ. 2003. *Proteomics* **3**:1567–1596.
6. Marengo E, et al. 2005. *Proteomics* **5**:654–666.
7. Appel RD, et al. 1997. *Electrophoresis* **18**:2735–2748.
8. Rogers M, Graham J, and Tonge RP. 2003. *Proteomics* **3**:879–886.
9. Rogers M, Graham J, and Tonge RP. 2003. *Proteomics* **3**:887–896.
10. Pleissner KP, et al. 1999. *Electrophoresis* **20**:755–765.
11. Aittokallio T, et al. 2005. *J Chromatogr B Analyt Technol Biomed Life Sci* **815**:25–37.
12. Veaser S, Dunn MJ, and Yang GZ. 2001. *Proteomics* **1**:856–870.
13. Zhou S, et al. 2005. *Proteomics* **5**:2739–2747.
14. Chang J, et al. 2004. *J Proteome Res* **3**:1210–1218.
15. Alban A, et al. 2003. *Proteomics* **3**:36–44.
16. Wheelock AM, et al. 2006. *Proteomics* **6**:1385–1398.
17. Wheelock AM and Buckpitt AR. 2005. *Electrophoresis* **26**:4508–4520.
18. Levanen B and Wheelock AM. 2008. Troubleshooting image analysis in 2-dimensional electrophoresis. In: *Two Dimensional Electrophoresis Protocols in Molecular Biology* (Sheehan D and Tyther R, eds.), Humana Press, Totowa, NJ, On press.

19. Miura K. 2001. *Electrophoresis* **22**:801–813.
20. Scrivener E, et al. 2005. *Proteomics* **5**:4354–4366.
21. Back P, Bengtsson S, and James P. 2005. *J Proteome Res* **4**:1511–1515.
22. Scheipers U, et al. 2005. *Ultrason Imaging* **27**:181–198.
23. Shapiro SS and Wilk MB. 1965. *Biometrika*: 591.
24. Karp NA and Lilley KS. 2005. *Proteomics* **5**:3105–3115.
25. Kreil DP, Karp NA, and Lilley KS. 2004. *Bioinformatics* **20**:2026–2034.
26. Gustafsson JS, et al. 2004. *Proteomics* **4**:3791–3799.
27. Karp NA, Griffin JL, and Lilley KS. 2005. *Proteomics* **5**:81–90.
28. Engkilde K, Jacobsen S, and Sondergaard I. 2007. *Methods Mol Biol* **355**:195–210.
29. Brereton RG. 2003. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley, West Sussex.
30. Molloy MP, et al. 2003. *Proteomics* **3**:1912–1919.
31. Corzett TH, et al. 2006. *J Proteome Res*.2006 **5**:2611–2619.
32. McKie A, Vyse A, and Maple C. 2002. *Lancet Infect Dis* **2**:18–24.
33. Handl HL and Gillies RJ. 2005. *Life Sci* **77**:361–371.
34. Tomita M. 2001. *Trends Biotechnol* **19**:205–210.

BIOINFORMATICS IN MS-BASED PROTEOMICS

Jacques Colinge

9.1 INTRODUCTION

This chapter aims at introducing fundamental knowledge in proteomic MS data processing. The increase in sensitivity, speed, and availability of high-performance proteomics technologies give researchers the possibility to generate large amounts of data. To provide useful scientific insight, these data must be properly processed though careful manual validation of each result is no longer possible at this scale. Consequently, software development, databases, and mathematical modeling are playing an important role in today data interpretation. Beside a general presentation of existing tools and well-established methods, we introduce basic concepts of data processing to give the readers the opportunity to learn about the tools underlying mechanisms. We consider top-down proteomics (analysis of enzymatic peptides) and MS/MS only (see Chapter 3) for a discussion of PMF, and we refer the reader to Chapters 3 and 27 for intact protein analyses.

9.2 DATABASE SEARCHING

Basic Notions

A natural way to identify proteins is by comparing experimental MS/MS data with protein sequences found in a database. We consider that the MS/MS data are already available as mass or peak lists—that is, as a sequence of experimental peptide masses and intensities accompanied by fragment masses and intensities as follows:

```
Peptide #1 mass, intensity
fragment mass, intensity
...
fragment mass, intensity
Peptide #2 mass intensity
fragment mass, intensity
...
fragment mass, intensity
Peptide #3 ...
```

For commodity, we refer to a mass list as a spectrum indifferently. The principle of database searching is the following: Each database protein sequence is digested theoretically (e.g., trypsin cleaves after lysine or arginine unless they are followed by proline); the experimental data are searched with these theoretical peptides masses; if one experimental peptide mass is found, then a theoretical fragmentation spectrum is computed from the peptide sequence and compared with the experimental fragmentation data; a scoring function determines the correlation between experimental and theoretical masses. The theoretical spectrum is the set of all possible fragment masses for a given instrument type; these masses can be computed by applying simple rules to the peptide sequences. The search ends by reporting every matched peptide with its highest score and by grouping the peptide matches by proteins (see Figure 9.1). Obviously, the quality of the scoring function plays an important role.

In practice, an MS instrument does not measure masses but measures m/z ; and depending on the instrument resolution, the charge z may be unknown. In this case the search algorithm must consider all possible charges within a certain range ($z = 1, 2, 3, 4$) for an ESI instrument typically, and each experimental m/z is considered as several possible masses.

Protein modifications, if present, result in peptide modifications after digestion, which change amino acid masses and hence peptide and fragment masses. To search for modified peptides, it is necessary to modify mass computations. Fixed modification (e.g., carboxyamidomethyl cysteines), does not really require additional work, because a mass is substituted for another mass. On the contrary, variable modifications cause additional computations. For instance, we may want to consider Met oxidation as a variable; and, consequently, a peptide ATMIQWMK would yield three peptide masses (0, 1, or 2 oxidations) and four theoretical fragmentation spectra (fragments masses are different, depending on which methionine is oxidized).

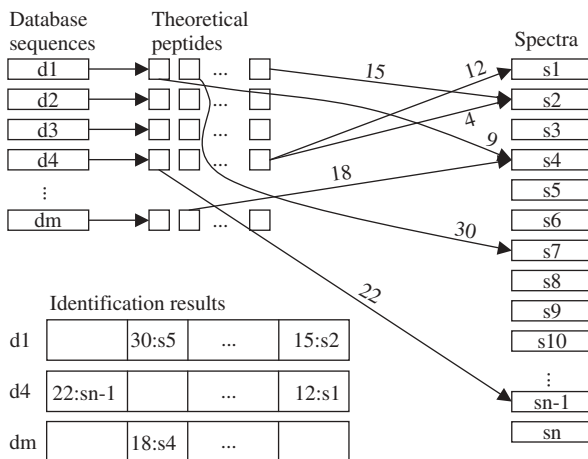


FIGURE 9.1. Scheme of a database search. The database sequences are digested theoretically, and each peptide is matched with 0, 1, or more spectra. Each match is given a score (number along the arrows). Most of the database sequences have no matching peptide, and many spectra are not matched. It is possible that a peptide matches several spectra (last peptide of d4) or a spectrum matches several peptides (s4). The peptide with the highest score is retained for each spectrum, and the spectrum yielding the highest score is retained for each peptide. The retained matches are grouped by database sequence and reported at the end of the search (lower left).

The last stage of the database search algorithm is the grouping of peptide identifications in protein identifications. Although straightforward in principle, this final step requires some care because it is possible that distinct proteins share some peptides. Depending on which peptides were identified, it may not be possible to distinguish one protein from another one on the basis of the detected peptides [1].

Databases for Plant Proteomics

To identify MS data by searching a database requires that the latter contain the protein sequence. In addition, it is desirable that the database provide reliable annotation regarding the protein function. One option is to choose an exhaustive database to identify as many spectra as possible. In this case, NCBI nr (<http://www.ncbi.nlm.nih.gov/>) or UniProt (<http://www.ebi.ac.uk/>) are possible choices. Alternatively, some plant genome sequencing projects provide almost complete coverage (<http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html>). Another option is a highly reliable database such as UniprotKB/Swiss-Prot (<http://www.ebi.ac.uk/>), which contains less redundant sequences, has less sequencing errors, and provides more functional annotations. An interesting compromise is the international protein index (IPI, <http://www.ebi.ac.uk/IPI/IPIhelp.html>), which integrates several high-quality databases. Unfortunately, *Ara-bidopsis* is the only plant covered by IPI at the time of writing. To obtain (long) lists of proteins after database search is not sufficient, and data integration is necessary

to organize results efficiently. Commercial and open-source packages offer a first level of integration by organizing identification results in a database and by making global queries and sample comparisons possible. A second level of data integration involves existing knowledge about proteins. Some existing packages comprise such functionalities, albeit with various degrees of sophistication and flexibility.

In complement to classical functional annotations, researchers use GO and gene/protein interaction databases in combination with quantitative information frequently. A detailed discussion of this topic is beyond the scope of this chapter, but it is important to realize that data integration should include these aspects as well, which is facilitated by well-annotated databases. Dedicated public projects focus on “beyond functional annotation” data integration—for example, ProteinHarvester (<http://harvester.embl.de/>), Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>), and iHop (<http://www.ihopnet.org/UniPub/iHOP/>).

Existing Tools

For several years the MS database search market has been dominated by two commercial products: Mascot [2], <http://www.matrixscience.com/>, and Sequest [3]. This quasi-monopole is now reduced, because new commercial and open-source tools offer competitive alternatives. We mention three such competitors: Phenyx [4], <http://www.phenyx-ms.com/>, OMSSA [5], and X!Tandem [6].

Hypothesis Testing

We have introduced the general algorithm of database searching. Now we discuss how this problem can be formalized to develop efficient solutions and to determine performance. Although the user interface, the search speed, and supported functionalities are very important elements in the choice of a database search tool, performance is governed by the quality of the scoring function. The appropriate framework to think about the problem of identifying MS data is the standard problem of hypothesis testing. In a hypothesis test problem we have to choose between two hypotheses, the null and the alternative hypotheses. As we compare an experimental spectrum with all the peptides having a compatible mass found in the database, we have to decide for each whether it is the correct one or not. The null hypothesis (H_0) is that it is not correct—or random—and the alternative hypothesis (H_1) is that it is correct. The decision between the two hypotheses is made on the basis of a number named a statistic. In our problem the statistic is the score returned by the scoring function. The quality of the scoring function influences the success and error rates of the decision process. To go further, we need to be precise how the decision is made and to define the possible types of errors.

To decide for H_1 means that the score is sufficiently high such that it seems very unlikely to come from a random match (H_0). Practically, we set a score threshold, and the peptide matches that are above the score threshold are named TP if they are correct and FP if they are random. Similarly, the matches under the score threshold are named true negatives (TN) if they are random and false negatives (FN) if they are

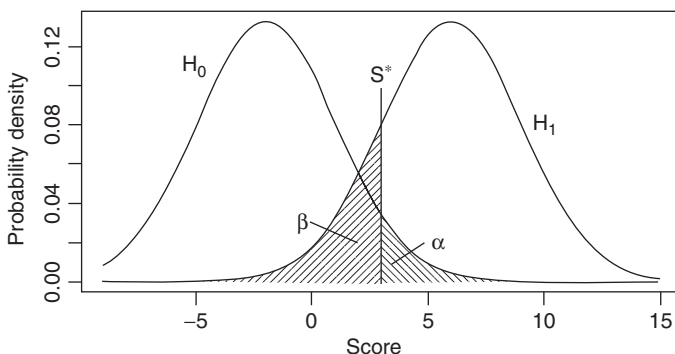


FIGURE 9.2. Hypothesis test. The value of the scoring function follows the H_0 distribution if the match is random and the H_1 distribution if the match is correct. In the figure we assume that both are Gaussian, but it is not always the case. If we set a score threshold s^* , then we find FP and FN rates α and β as indicated in the figure. In this example we chose $s^* = 2.93$ such that $\alpha = 0.05$ and $\beta = 0.15$.

correct. The FP rate is denoted α , and the FN rate is denoted β (see Figure 9.2). If we assume that we know the probability distribution of the score under H_0 , we can set the score threshold to achieve a given error rate.

There exist methods to learn the random score distribution during database search or *a priori*, and hence it is possible to warrant a given FP rate α . To determine the correct score distribution is not possible usually at the time of the search. To compare scoring functions, it is common practice to use a set of spectra whose peptide identifications are known. We search the spectra against a database that contains the matching peptide sequences (H_1) and against a database that does not (H_0). For a fixed score threshold we can deduce from the first database the TP and FN rates, $1 - \beta$ and β respectively, and from the second database the FP and TN rates, α and $1 - \alpha$ respectively. To capture the scoring function performance globally, the previous rates are determined by varying the threshold and a curve is obtained, the so-called ROC curve (see Figure 9.3). Since mass lists can contain more or less masses, some spectra will match wrong peptide sequences with relatively large or low scores. This difficulty is normally circumvented by computing a normalized score (z -score, subtraction of the mean, and division by the standard deviation) and/or a p -value—that is, the probability to observe a score as high or higher by random chance (H_0). When a z -score or a p -value is computed, they normally replace the score, which becomes of secondary interest though it still conditions the search tool performance.

It is essential to realize that (i) the scoring function is just a formula or an algorithm (i.e., it is just a model), and it cannot be perfect, and (ii) the computation of the z -score or the p -value involves another model (e.g., we assume normality of the random scores). Consequently, the scores and the p -values returned by a search engine must always be regarded with a minimum degree of criticism, a situation that is analogous to what happens in sequence homology.

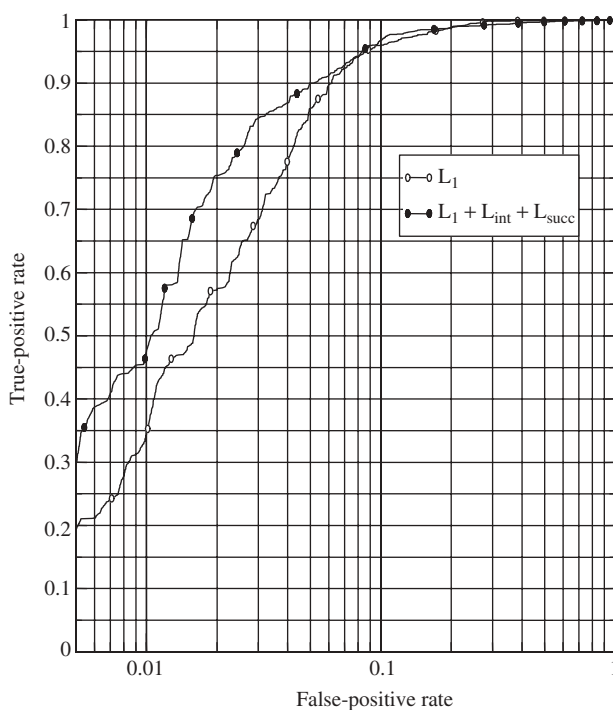


FIGURE 9.3. ROC curves. Comparison of two scoring functions performance when searching ion trap MS/MS spectra against the entire Swiss-Prot database. $L_1 + L_{\text{int}} + L_{\text{succ}}$ includes the model of ion intensities and successive fragment matches, whereas L_1 includes the model of independent fragment matches only. Scoring functions L_1 and $L_1 + L_{\text{int}} + L_{\text{succ}}$ are designed as likelihood ratios (see text).

Examples of Scoring Functions

Let us first review well-accepted characteristics of correct peptide matches: As many as possible experimental masses should be matched, especially the most intense peaks; sequences of successive fragments (increments of one amino acids) should be observed; certain ion types generate more intense signals (e.g., y and b), than other ones (e.g., a- and b-H₂O), and therefore they should contribute more to the score. More complicated criteria can be applied, but we ignore them here. There exist scoring functions based on statistical models and scoring functions that are purely heuristic. Since the fragmentation spectrum is obtained as the sum of signals generated by many copies of the selected peptide, to use a statistical approach makes a lot of sense. It is the main approach today, although there are heuristic scoring functions—old and new—that perform relatively well.

The Sequest approach to the scoring problem comprises two heuristic scoring functions: s_n , a fast heuristic function that combines some of the characteristics above, and X_{corr} , a more elaborated function that is used to re-score the 200 best peptides

found by s_n for each spectrum. The function s_n is given by a simple formula that sums the intensity of the matched peaks, multiplied by the number of matched masses, and so on. A correct factor is applied for observed sequences of consecutive fragment matches [7]. X_{corr} is founded on a cross-correlation ($a * e$) between the experimental mass list e and an artificial spectrum a deduced from the theoretical mass list (a peak intensities depend on the ion types):

$$X_{\text{corr}} = \frac{(a * e)(0) - E\{(a * e)(t); t \in [-75; 75]\}}{(a * a)(0) - E\{(a * a)(t); t \in [-75; 75]\}},$$

where $(a * e)(t)$ is now a cross-correlation with delay t , which introduces a shift in the list of masses; that is, a mass m from a is compared with a mass $m + t$ in e . X_{corr} formula can be explained as follows: $(a * e)(0)$ is the true cross-correlation between a and e , and it is corrected by the mean (operator E) of a similar cross-correlation when a shift between -75 and 75 Da is applied. The role of this mean is to correct for typical random match values. The division by a similar expression, where a is cross-correlated with itself, is a normalization factor because it represents the best possible score. The mean correction and the normalization are Sequest solutions to the problem we mentioned regarding dense mass spectra that give biased high scores and sparse spectra that give biased low scores. Sequest also includes experimental spectrum intensities normalization before it is used in the above computation.

We see that the Sequest approach includes all the criteria for correct matches, although their combination is purely heuristic. Several authors developed postprocessors to model all the quantities output by Sequest in machine learning models to embed Sequest into a limited statistical framework. Sequest does not identify proteins directly but instead identifies peptides; additional tools are used to group peptides in proteins such as DTASelect.

The Mascot scoring function was never disclosed at the time of writing, and hence we limit its presentation to some general statements. As Sequest does, some spectrum preprocessing is performed, followed by the normal match between theoretical and experimental masses. On the basis of this match, Mascot selects a limited number of ion types—usually two—where most fragment matches are found to compute its score. For instance, Mascot might compute its score based on y and $y\text{-NH}_3$ fragments and ignore all the matches found in the other types. This procedure is intended to improve score robustness obviously, though significant information is ignored. The returned so-called ion score is the negative logarithm of a p -value computed from a statistical model. Mascot defines the protein score as the sum of the peptide scores.

Since the design of the scoring function is free, it is natural to try to create the best possible one. Unfortunately, this task is not realistic, and high-performance functions can be imagined only. Nonetheless, hypothesis testing theory provides one useful guideline. Let M denotes the observed match when comparing a candidate peptide sequence and an MS/MS spectrum. Neyman–Pearsons’s Lemma—with some adaptation—says that, provided the exact probabilities to observe M under the null

and alternative hypotheses can be computed, then the optimal score is the likelihood ratio:

$$L = \frac{P(M|H_1)}{P(M|H_0)}$$

The probabilities above can be approximated by means of statistical or machine learning, but then L is no longer optimal. In practice, an approximation of this ratio is often an excellent scoring function, and it is not a surprise if most—but not all!—of the recent functions follow this principle. Likelihood ratios are present in many bioinformatics tools such as sequence alignment substitution matrices (PAM, BLOSUM) and search tools (BLAST, Pfam), as well as in tests aimed at detecting regulated genes in microarrays.

As an example of a likelihood ratio-based score, we describe the model used by Phenyx. This model is an extension of a simple model [8], which rewards each matched fragment separately. The principle of the Dancik et al. model is to consider each fragment match as an independent event, which is scored by a likelihood ratio:

$$L_s = P(\text{"fragment match"} | \text{"ion type"} = s, H_1) / \\ \times P(\text{"fragment match"} | \text{"ion type"} = s, H_0)$$

We see that the probabilities are dependent on ion type, and they can be learned easily from a set of correct and random matches. For instance, in the case of ion trap doubly charged tryptic peptides, we find $L_b = 0.57/0.13 = 4.38$, $L_y = 0.61/0.11 = 5.55$, and $L_{y++} = 0.17/0.09 = 1.89$, and for triply charged peptides we have $L_b = 0.35/0.10 = 3.50$, $L_y = 0.40/0.10 = 4.00$, and $L_{y++} = 0.39/0.16 = 2.44$, which illustrates the different amount of information provided by each type of ion and the dependence on the peptide charge state. Because we assume the fragment matches to be independent events, we obtain the peptide score L_1 by multiplying the fragment scores.

This first model is improved by adding models for consecutive fragment matches, peak intensities, and amino acid composition [4], see Figure 9.3 for a comparison of the Dancik et al. model with the same model combined with models for peak intensities and contiguous fragment matches. As explained, random scores distributions are learned during database search, and z -scores and p -values are obtained accordingly. Phenyx computes protein scores as Mascot does.

Performance Achieved

To discuss performance is always a difficult exercise because performance is data-dependent and, pure TP/FP performance considerations must be weighed against tool functionalities. We believe to have demonstrated superior performance of advanced statistics-founded scoring functions (Phenyx) compared to heuristic or simpler functions [4], but, admittedly, this sort of advantage diminishes with excellent data quality and mass accuracy (~ 10 ppm on peptide masses). Since these two conditions are not always satisfied, scoring functions are not anonymous components. It is also a common belief that Mascot achieves better results than Sequest with TOF data, and Sequest

performs better than Mascot with ion trap data. OMSSA and X!Tandem performance is usually reported comparable to Mascot and Sequest. X!Tandem was designed to be very fast, and it comes with a concept of data representation and sharing (the global proteome machine). We also stress that no scoring function is general enough and exhaustivity can be achieved by using more than one search tool only [9].

9.3 PEPTIDE *DE NOVO* SEQUENCING

Introduction

Database search does not work if the correct peptide (protein) sequence is not in the database [10] or if the peptide is modified in an unanticipated manner [11]. The first scenario is often valid for organisms whose genomes are not sequenced completely such as numerous plants. Instead of searching spectra against a sequence database, it is possible to try to infer complete or partial peptide sequences from the MS/MS spectra directly. This difficult problem is named peptide *de novo* sequencing in general, but it is called sequence tag prediction when we refer to partial predictions specifically. A third potential application to *de novo* sequencing is to reduce compute time when searching extremely large databases. The idea is to obtain reliable partial predictions first and then to filter the database based on the latter before final search by the standard algorithm of Figure 9.1.

Current Approaches

Successful *de novo* sequencing algorithms are more diverse than database searching algorithms. A common initial step is to reduce the experimental mass list size by trying to remove noisy peaks. Additional preprocessing steps can be considered like normalizing intensities or trying to improve experimental peptide mass accuracy [8]. As a second step, there are algorithms that try to predict the ion type corresponding to each experimental mass [12]. As a matter of fact, if we know which masses in a spectrum are b or y ions, then we can read the sequence directly (see Figure 9.4). In practice, this step cannot be perfect, and numerous additional masses are predicted as b or y and hence the problem difficulty is reduced only.

After this optional second step, four main types of methods are employed. The most common is the spectrum graph: A graph is built by creating one node per experimental mass plus one for the entire peptide and one empty node. The nodes, whose masses differ by one amino acid mass, are linked and the edges labeled with the amino acids (see Figure 9.4). In the graph, the correct peptide sequence is assumed to appear as a path from the empty node to the entire peptide node. Due to possibly missing b or y fragment masses, which would disrupt the correct path, experimental masses are complemented before the graph is built: $m \rightarrow \text{mass}(P) + 2 - m$, where $\text{mass}(P)$ the mass of the peptide P . This complementation changes the mass of a b ion into the complementary y ion and vice versa. The spectrum graph is then exploited either empirically or—more frequently—by scoring each edge and by searching for the longest path [13]. To score edges requires scoring functions similar to the

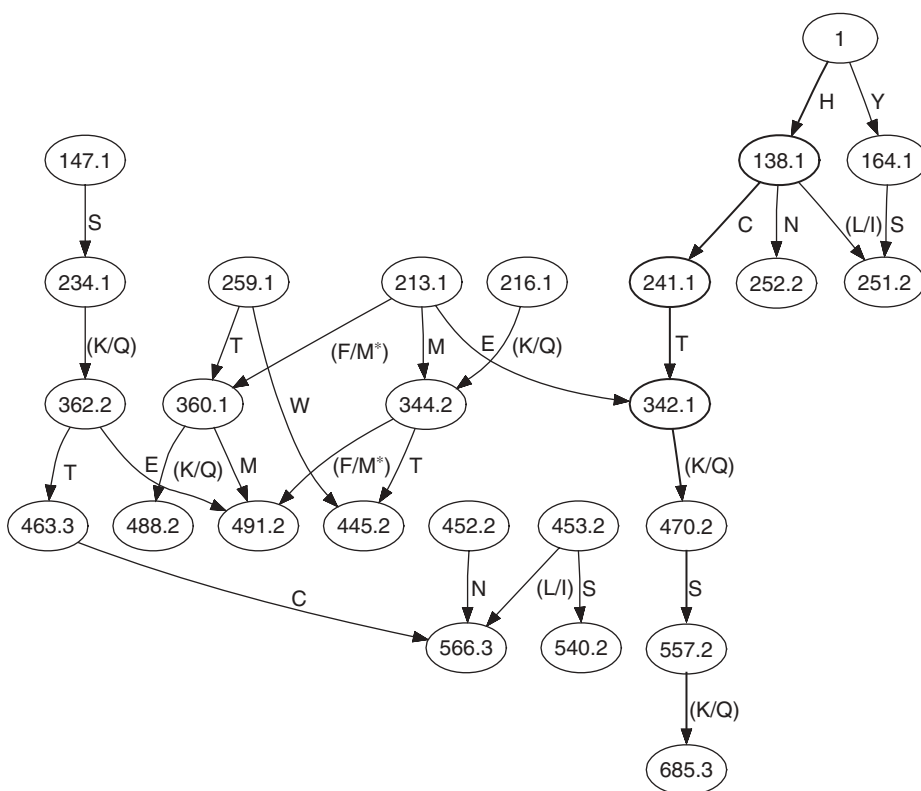


FIGURE 9.4. Example of a spectrum graph. We used the theoretical spectrum of the peptide HCTQSK (b and y ions plus a few neutral losses), and we removed some b and y masses to obtain a very simple spectrum. The correct path and the b masses obtained by complementing the y masses are in bold. Without complementation, only a sequence tag was possible to compute (note the partial reversed sequence SQTC created by the noncomplemented y masses). A real spectrum with noisy peaks would yield a more complex graph with many alternative paths from 1 to 685.3.

ones used for database searching, and the likelihood ratio-based scores are prevalent in *de novo* sequencing. In general, the longest-path problem is intractable, but spectrum graphs have properties that make fast dynamic programming algorithms possible [14].

Given a peptide scoring function as we used for database searching, the *de novo* sequencing problem can be formulated as an optimization problem over the space of all the possible peptide sequences: Find the highest-scoring sequence [15]. Markov chain Monte Carlo methods (e.g., genetic algorithms), are used to perform the optimization. The principle of the genetic algorithm is to generate a population of random peptide sequences that have a mass close to the experimental peptide mass. These sequences are then modified by changing amino acids (mutation) or by combining two sequences

from the population (cross-over). Each new set of modified sequences constitutes a new generation, from which only the sequences with the highest scores are taken into account to generate the next generation. After a certain number of generations, it is possible to obtain high-scoring sequences.

One proposed algorithm sees the MS/MS spectrum peaks as generated by a sequential process and therefore apply an HMM to model peak intensities [16]. The model follows an amino acid sequence to generate intensities and thus produce a sequence for the optimal state path. Finally, it is possible to try to build partial sequences from each terminus of the peptide and to join them. This can be done empirically, but an optimal algorithm has been proposed that computes the optimal peptide sequence for a certain class of scoring functions by dynamic programming [17].

Existing Tools

Due to its challenging nature, the *de novo* peptide sequencing problem attracted many researchers and there are numerous tools available. A partial list comprising well-established algorithms and some challengers is: PepNovo [13], PEAKS [17], Lutefisk [18], NovoHMM [16], InsPecT [19], and EigenMS [12].

Performance Achieved

Large-scale algorithm comparisons and validations are missing, but what is currently reported is that the best algorithms are PEAKS and PepNovo [20], with an advantage for PEAKS with TOF data and for PepNovo with ion trap data. In every case, it is essential to mention that the typical performance achieved is around 85–90% correct partial sequence predictions if we accept sequence tags of length 3–5 and we work with good-quality spectra only. Moreover, this is the performance for ESI doubly charged peptides; otherwise the performance is worse. This performance is sufficient to obtain satisfying results when several peptides are available for a protein, but it is still limited intrinsically and *de novo* results should be considered with more care than database search results.

Genome and Exhaustive Searches

To search MS/MS spectra against a genome sequence after translation in the six reading frames gives the opportunity to confirm and correct genome annotations by using experimental data [21]. This approach is integrated in several large-scale proteomics projects [22] now, and its potential has been validated [23]. Proteomics MS analyses remain with a large proportion of unexplained good-quality spectra, roughly 60–70%. To attempt to explain as many as possible spectra, it is customary to cascade identification procedures from the most reliable to the less reliable: Spectra are searched against small but reliable databases, they are then searched against larger and more error prone databases, and *de novo* predictions are computed finally. At each step, a spectrum that is matched reliably is removed.

Other Problems

Proteomics technologies generate large amounts of data of different types. We briefly mention two other data analysis problems of interest currently: (1) LC chromatogram alignments and comparisons and (2) MS profiles. The 2D gels are widely used for separating proteins. Direct sample comparisons can be performed on the basis of the gel images in addition to MS or to limit MS to potentially interesting spots only (see Chapter 8). Similar techniques have been developed to work with LC–MS ion chromatograms. The intensity of each peptide ion is followed along the time axis, thus generating a 2D display m/z versus time. By aligning and comparing such maps as obtained from several samples, it is possible to identify variable patterns of potential interest.

Samples can be compared by direct MS analysis, generally with a MALDI instrument, and the spectra acquired not only contain peptides and small protein masses but other analytes masses as well. This type of analysis is often referred to as metabolomics (see Chapter 7 for a related topic). Before comparison, the spectra must be aligned because slight variations in the masses have to be corrected to recognize identical molecules. After this procedure, masses are associated with an expression profile given by the peak intensities, and multivariate statistics methods can be applied to detect significant variations.

9.4 CONCLUSIONS AND FIVE-YEAR VIEWPOINT

In recent years a lot of attention has been devoted to improve protein identifications reliability, to develop quantitative methods with associated data processing, and to integrate data. These efforts are not yet fully completed, but they already changed proteomics profoundly. Over the next years, these efforts will all continue, but the main focus will be on data integration certainly. It is a complicated problem covering the management of protein identifications at the level of individual samples, sample comparisons, combination with mRNA microarray data, protein modification data (phosphorylation mainly), and integration with all kinds of existing biological knowledge. All these techniques already exist, but their efficient combination and distribution in every proteomics laboratory is far from being achieved.

In addition to the above-mentioned topics, systems biology type of studies will intensify and necessitate always more network analysis and predictive methods (see Chapter 46). Therefore the application of proteomics as a component in more global studies will increase, thereby stressing even more the need for complete heterogeneous data integration. The development of more reliable protein identification procedures and the aforementioned developments attracted “classical bioinformatics” specialists in proteomics and MS data processing. The continuation of these efforts and advanced data integration will merge the previously relatively isolated proteomics data analysis community with mainstream bioinformatics research as the shared techniques always become more important (sequence analysis, data integration, graphic display, expression profiles analysis, etc.).

REFERENCES

1. Nesvizhskii AI and Aebersold R. 2005. *Mol Cell Proteomics* **4**:1419–1440.
2. Perkins DN, et al. 1999. *Electrophoresis* **20**:3551–3567.
3. Eng JK, McCormack AJ, and Yates JR 3rd. 1994. *J Am Soc Mass Spectrom* **5**:976–989.
4. Colinge J, et al. 2004. *Proteomics* **4**:1977–1984.
5. Geer LY, et al. 2004. *J Proteome Res* **3**:958–964.
6. Craig R, Cortens JP, and Beavis RC. 2004. *J Proteome Res* **3**:1234–1242.
7. Yates JR 3rd and Eng JK. 1996. Identification of nucleotides, amino acids, or carbohydrates by mass spectrometry. United States Patent 6,017,693.
8. Dancik V, et al. 1999. *J Comp Biol* **6**:327–342.
9. Heller M, et al. 2005. *J Proteome Res* **4**:2273–2282.
10. Sunyaev S, et al. 2003. *Anal Chem* **75**:1307–1315.
11. Searle BC, et al. 2004. *Anal Chem* **76**:2220–2230.
12. Bern M and Goldberg D. 2006. *J Comput Biol* **13**:364–378.
13. Frank A and Pevzner P. 2005. *Anal Chem* **77**:964–973.
14. Chen T, et al. 2001. *J Comput Biol* **8**:325–337.
15. Heredia-Langner A, et al. 2004. *Bioinformatics* **20**:2296–2304.
16. Fischer B, et al. 2005. *Anal Chem* **77**:7265–7273.
17. Ma B, et al. 2003. *Rapid Commun Mass Spectrom* **17**:2337–2342.
18. Taylor JA and Johnson RS. 2001. *Anal Chem* **73**:2594–2604.
19. Tanner S, et al. 2005. *Anal Chem* **77**:4626–4639.
20. Pevtsov S, et al. 2006. *J Proteome Res* **5**:3018–3028.
21. Colinge J, et al. 2005. *J Proteome Res* **4**:167–174.
22. Desiere F, et al. 2006. *Nucleic Acids Res* **34**:655.
23. Kuster B, et al. 2001. *Proteomics* **1**:641–650.

PART III

EXPRESSION PROTEOMICS

AN OVERVIEW OF THE *ARABIDOPSIS* PROTEOME

Jacques Bourguignon and Michel Jaquinod

10.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

***A. thaliana*: A Model Plant for Functional Genomics**

A. thaliana has been adopted as a model organism by plant biologists because of the small size of its genome (125 MBp), its short generation time, its high transformation efficiency, and the large panel of mutants generated. While the world turned its attention to the project of the human genome, the first complete genome published in December 2000 in the *Nature* review was none other than *Arabidopsis*, the first flowering plant to be sequenced [1] (Table 10.1). Mitochondrial and plastid genomes had been sequenced shortly before [2, 3]. A direct comparison between the genomic sequence and data banks of ESTs suggests that the *Arabidopsis* genome contains roughly 26,000 genes. About 65% of the predicted proteins have significant homology with proteins from other organisms, and functional assignments were established [4]. The remaining 35% of the predicted proteins can be expected to contribute to plant-specific processes. Although alternative splicing is thought to be less prevalent in plants than it is in other higher eukaryotes, recent computing analysis in *Arabidopsis* suggest that 1333 genes have alternatively spliced isoforms generating

TABLE 10.1. Useful Internet Resources for Proteomics Studies

<i>Arabidopsis thaliana</i> Genome	
TIGR chromosomes I, II, and III	http://www.tigr.org/tdb/at/atgenome/atgenome.html
<i>Arabidopsis</i> genome analysis chromosomes IV and V	http://nucleus.cshl.org/protarab/
The Kazusa <i>Arabidopsis</i> opening site III and V	http://www.kazusa.or.jp/kaos/
Genome SPP chromosome I	http://med.stanford.edu/sgtc/research/arab.html
GENOSCOPE chromosome III	http://www.genoscope.cns.fr
General Databases	
The European <i>Arabidopsis</i> Stock Centre (NASC)	http://arabidopsis.info/
The <i>Arabidopsis</i> Information Resource (TAIR)	http://www.arabidopsis.org/index.jsp
The TIGR <i>Arabidopsis thaliana</i> Genome Annotation	http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml
<i>Arabidopsis</i> Seed Proteome	http://seed.proteome.free.fr/
Proteomic Data <i>Arabidopsis thaliana</i> (GABI)	http://gabi.rzpd.de/
Plant Protein Membrane Databases	
<i>Arabidopsis</i> Membrane Protein Library (AMPL)	http://www.cbs.umn.edu/arabidopsis/
Aramemnon	http://aramemnon.botanik.uni-koeln.de/
Transport Classification Database (TCDB)	http://www.tcdb.org/index.php
Transport Database	http://www.membranetransport.org/
Specialized Databases	
MEROPS the Peptidases Database	http://merops.sanger.ac.uk/
MIPS <i>Arabidopsis thaliana</i> Database	http://mips.gsf.de/proj/plant/jsf/athal/index.jsp
Functional Genomics of Plant Phosphorylation (PlantsP)	http://plantsp.genomics.purdue.edu/
Database of <i>Arabidopsis</i> Transcription Factors (DATF)	http://datf.cbi.pku.edu.cn/
<i>Arabidopsis</i> Lipid Gene Database	http://www.plantbiology.msu.edu/lipids/genesurvey/
Subproteome Databases	
<i>Arabidopsis</i> Proteins from Plant Peroxisome (araperox)	http://www.araperox.uni-goettingen.de/
The Plastid Proteome Database (PPDB)	http://ppdb.tc.cornell.edu/
Plastid Proteome Database (plprot)	http://www.plprot.ethz.ch/
Chloroplast Genome dB	http://chloroplast.cbio.psu.edu/
<i>Arabidopsis</i> Mitochondrial Proteome Project (AMPP)	http://www.gartenbau.uni-hannover.de/genetik/AMPP

TABLE 10.1. (Continued)

<i>Arabidopsis</i> Nucleolar Protein Database (AtNoPDB)	http://bioinf.scri.sari.ac.uk/cgi-bin/atnopdb/home
SubCellular Proteomic Database (SUBA)	http://www.plantenergy.uwa.edu.au/applications/suba/
Cell Wall Navigator (CWN)	http://bioweb.ucr.edu/Cellwall/index.pl
Protein Interaction Databases	
Biomolecular Object Network Databank (BOND)	http://bond.unleashedinformatics.com/Action?
Molecular Interaction Data (IntAct)	http://www.ebi.ac.uk/intact
Molecular Interaction Database (MINT)	http://mint.bio.uniroma2.it/mint/
Database of Interacting Proteins (DIP)	http://dip.doe-mbi.ucla.edu/
Dedicate Tools for Prediction of Organelle Targeting Sequences and Localization	
Genoplante Predotar (ER, mitochondrial or plastid transit peptides)	http://urgi.infobiogen.fr/predotar/
SignalP (signal peptide cleavage sites)	http://www.cbs.dtu.dk/services/SignalP/
ChloroP (chloroplast transit peptides)	http://www.cbs.dtu.dk/services/ChloroP
Mitopred (predicting mitochondrial proteins)	http://bioinformatics.albany.edu/~mitopred/
WoLFPSORT (protein subcellular localization prediction)	http://wolfpsort.seq.cbrc.jp/
PredictNLS (prediction and analysis of nuclear localization signals)	http://cubic.bioc.columbia.edu/services/predictNLS/
PredSL (prediction of subcellular location)	http://bioinformatics.biol.uoa.gr/PredSL
PrediSi (prediction of signal peptides)	http://www.predisi.de
iPSORT (predictor for N-terminal sorting signals)	http://hc.ims.u-tokyo.ac.jp/iPSORT/
Loctree (predicting the subcellular localization of proteins)	http://cubic.bioc.columbia.edu/services/loctree/

over 2600 proteins in total [5] (Table 10.1). From a functional point of view, it can be pointed out that only 10% of *Arabidopsis* genes have been characterized experimentally, more than 35% of proteins are annotated as “expressed protein” in the *Arabidopsis* genome, another 10% are annotated as “hypothetical protein,” and 2.5% are referred to as “unknown protein.” Proteomics is essential for understanding the function of uncharacterized proteins, specific to plants or not, and to provide novel insights into plant cell biology. The recent important development of analytical methods based on MS has allowed both qualitative and quantitative analyses of proteins at the subpicomole level. The rapid identification of a large number of proteins within a given biological sample has led proteomics to become a major field of functional genomics in complement to the analysis of transcriptome and metabolome. Whereas

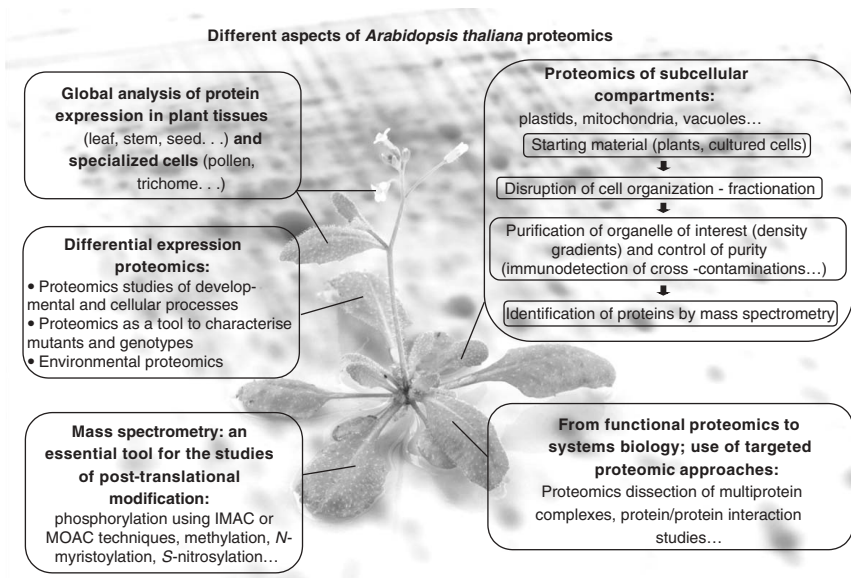


FIGURE 10.1. An overview of the different topics of the *A. thaliana* proteomic studies.

proteomics research is quite advanced in animals and yeast as well as *E. coli*, plant proteomics is only at the initial phase. In this chapter we present an overview of the *Arabidopsis* proteome studies, from the obtaining of protein repertoires to the presentation of various functional aspects of proteomics such as differential expression proteomics, PTMs and PPIs. An overview of the different topics is summarized in Figure 10.1.

10.2 METHODOLOGY AND STRATEGY

Classical proteomics strategies involve a separation step using 2-DGE separation followed by an identification one, which is usually based on MS identification via PMF or MS/MS. A large part of proteomics work on *Arabidopsis* has used this combined technology. However, gel-based proteomics strategies are rapidly being supplanted by methods that involve the use of in-solution protein extract digestion and peptide fractionation by liquid peptide separation via high-efficiency nanocolumn LC separation techniques directly linked to a mass spectrometer. In these methods, complex protein mixtures are reduced to peptides prior to separation, and the total peptide mixture is loaded onto a nanocolumn. MS methods can resolve thousands individual peptides at a time, but it is estimated that there are over 10,000 proteins in any given cell type. A proteome consisting of 10,000 proteins will generate approximately 350,000 peptides. Therefore, peptides have to be separated prior to mass spectral analysis, and their separation is the bottleneck of the proteomic approach because the number of peptides

easily exceeds the resolution capacity of front-end separation techniques. To counteract this difficulty, another way to reduce the peptide complexity prior to MS analysis consists of organelle isolation or targeted strategies such as affinity-based purification techniques.

10.3 EXPERIMENTAL RESULTS AND APPLICATIONS

Global Analysis of Protein Expression in Plant Tissues and Specialized Cells

The combination of an efficient method for the separation of protein or peptide mixtures with subsequent identification by MS can provide a repertoire and/or map of the major proteins present in plant tissues or cells. Despite limitations of the method, the use of 2-DGE is a mode of separation and visualization of complex protein mixtures. 2-DGE separate large numbers of proteins and continues to be popular. The capacity to display the protein content of a sample “as an image” in a simple manner, giving an approximation of the pI and MW of proteins, made this a method of choice in several analyses of *Arabidopsis* tissues. The first protein profiling investigation performed was presented in 1995 [6] with proteins extracted from leaf, stem, root, seed, and callus. Proteins were analyzed using 2-DGE; and a 2D gel reference map was established, revealing a total of 4763 protein spots from all these tissues. Proteins were characterized using Edman degradation of the N-terminal amino acid sequences, which led to only 57 identifications. Recently, a large-scale proteome analysis conducted with eight different *Arabidopsis* tissues (primary leaf, leaf, stem, silique, seedling, seed, root, or inflorescence) were carried out [7]. 2D gel reference maps from three selected tissues were established using two different extraction protocols (Table 10.1). A total of 6000 spots were analyzed by MALDI–TOF–MS. This work allowed the identification of 2943 proteins, products of only 663 different genes. It is noteworthy that other 2D gel reference maps were also established from specialized tissues such as seed [8, 9] (Table 10.1) or cells: mature pollen [10, 11], trichome [12], or cultured cells [13]. The establishment of these reference maps allowed the characterization of a set of proteins present in a cell, or a tissue at a given time and at a particular developmental stage, and constitutes a basis to explore proteome dynamics. We present below examples of studies that analyze proteome changes in response to a perturbation. Nevertheless, the problem of this approach, when characterizing a global proteome, is that the most abundant proteins present in the sample are repeatedly analyzed, and these abundant proteins prevent detection and identification of the low-abundance proteins. As an example, the most abundant protein in plants, the LSU of RuBisCO, was also the protein identified the most frequently in each of the 2D gels analyzed [7]. Alone, this protein represented a total of 366 spots, corresponding to 12.5% of all assigned spots. Low-abundance proteins are out of the scope of these global proteomics analyses, and specific strategies must be developed to analyze subproteomes; by reducing protein extract complexity, these allow study of rarer proteins.

Proteomics of Subcellular Compartments

The characterization of proteomes at different subcellular locations is of prime importance for a complete understanding of plant functions, biosynthetic, and signaling pathways. The use of subcellular fractionation permits a simplification of the proteome and provides a practical step toward the ultimate description of the entire proteome. The efficiency of proteomics technology relies essentially on the quality of the biological sample analyzed; thus sample preparation becomes the most critical step in subcellular compartment proteomics. Despite the fact that *Arabidopsis* was not the most appropriate plant on which to carry out subcellular fractionation, a number of procedures have been developed to enrich and purify organelles. The choice of starting material is important to prepare highly purified organelles. For example, the choice of leaf tissue to prepare chloroplasts seems obvious [14–16]. Mitochondria are better purified from cell suspensions [17–19]. The classical cell fractionation procedure usually consists of two major steps: disruption of the cellular organization and fractionation of the crude homogenate in order to purify the organelle of interest (Figure 10.1). The first step has to be controlled to avoid disruption of subcellular compartments. The second step generally consists of a series of differential centrifugations that enrich the targeted intact organelle and eliminate, as far as possible, contaminants from other compartments, followed by separation on density gradients (sucrose, Percoll, or Ficoll). Organelle purity is essential and can be estimated by calculating the enrichment factor for the organelle of interest and by estimation of the cross-contamination by other compartments. This can be monitored by immunodetection to follow the distribution of specific organelle-marker proteins and/or by measuring enzymatic-marker activities. To date, several studies have also used a prefractionation protocol to separate hydrophobic and hydrophilic components to gain insight into the subcellular functions of an organelle [15, 19–21]. Membranes, and proteins they contain, occupy a fundamental place in the complex organization of the cell because of their position at the interface between cells or between cell compartments. They control exchange of signals and solutes and allow the compartmentalization of specific biochemical pathways. The distribution of membrane proteins reported on AMPL database (Table 10.1) indicates that roughly 4750 *Arabidopsis* proteins contain two or more TMDs. In spite of the recent advances in techniques for the separation and analysis of membrane proteins, characterization of these proteins is still challenging. Nevertheless, strategies for large-scale identification of membrane proteins have been explored. Classical 2-DGE or BN-PAGE combined with the use of appropriate detergents can be very efficient for the analysis of hydrophobic membrane proteins [18, 22–24]. Nevertheless, a large part of the MS-based identification of *Arabidopsis* membrane proteins, from plastids, mitochondria and PM, was performed using (1D) SDS-PAGE fractionation combined with extraction methods: based on organic solvent extraction using a mixture of chloroform/methanol (C/M) [15, 19–21, 25], salt and/or alkaline treatments [15], detergent treatment such as sodium deoxycholate [26], or Triton X-114 phase partitioning [20, 21, 24, 27]. A method of in-solution proteolytic digestion, in the presence of methanol, has recently been used to identify very hydrophobic vacuolar proteins [28]. Databanks focused on plant membrane proteomes, compiling experimental data, topology, and

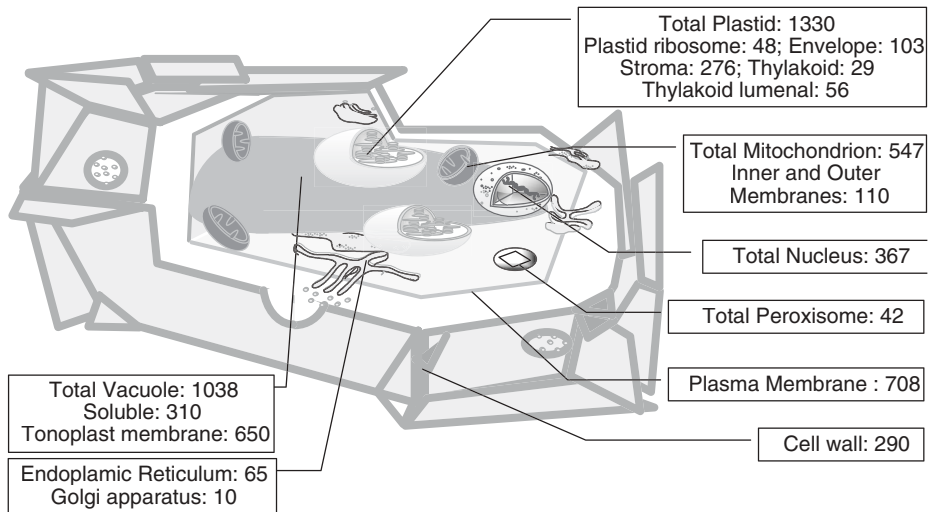


FIGURE 10.2. *A. thaliana* proteins identified by proteomics and their repartition in subcellular compartments (December 2006). See insert for color representation of this figure.

intracellular targeting predictions are available on the web (Table 10.1) [for review see reference 29].

Beyond the establishment of protein catalogues provided by high-throughput technologies, organelle proteomics studies are leading to major discoveries in plant biology and are helping to reshape our understanding of the functional organization of cells at the molecular level. A snapshot of MS-identified proteins isolated from different subcellular compartments of the *Arabidopsis* cell and related databases are given in Figure 10.2 and Table 10.1; selected examples, focused on plant particularities, will be overviewed below.

The cell wall most commonly found in plant cells is an extracellular structure surrounding the plasmalemma composed of carbohydrate polymers (celluloses, hemicelluloses, and pectins), proteins, and phenolic compounds. The cell wall is a dynamic structure that is essential for development and morphogenesis and is involved in cell division and differentiation, as well as in responses to biotic and abiotic stress [30]. Nevertheless, cell wall modification and expansion are not fully understood, and several groups have attempted to establish the *Arabidopsis* cell wall proteome and secretome (secreted proteome) [31–33], “loosely bound” cell wall proteome [34–36], or “tightly bound” proteome [37] (Table 10.1). The most complete study, based on a MudPIT approach, led to the identification of 792 unique proteins, 106 of which were predicted to have a TM span. Analysis of the proteomics data was completed using bioinformatics tools for the prediction of signal peptides for targeting to the secretory pathway, leading to the classification of 89 proteins as potential extracellular proteins, notably carbohydrases, peroxidases (POX), and proteases [37].

Nuclear proteins were isolated and analyzed for the first time in 2003 using 2-DGE and MALDI–TOF–MS [38]. Proteomic analyses led to the identification of

proteins implicated in a variety of cellular functions such as signaling, gene regulation, structure, translation, proteolysis, and various RNA-associated functions. Recently, the first proteomic analysis of nucleolus was carried out: 217 proteins were identified and 27 were expressed as GFP fusions [39]. An *Arabidopsis* nucleolar protein database (Table 10.1) was built, providing (a) information on the plant proteins identified to date with comparison to human and yeast proteins and (b) images of cellular localizations for over a third of the proteins [40]. Nevertheless, we are only at the beginning of nuclear protein characterization. Preliminary analyses of *Arabidopsis* DNA sequence, using prediction software (LOCtree), surprisingly suggests that over 35% of all nonmembrane proteins are nuclear and only about 20% are located in the cytosol [41].

Organelles with Double Membranes: Plastids and Mitochondria. Evolution of the eukaryotic cell is thought to have occurred by a process of endo-symbiosis. One piece of evidence that plastids and mitochondria arose by this process comes from the observation that these organelles are surrounded by double membranes. Since the chloroplast and mitochondrial *Arabidopsis* genomes contain approximately 79 and 57 genes, respectively [2, 3], most proteins are encoded by the nuclear genome and synthesized in the cytosol before being delivered to their final destination. These nuclear-encoded proteins usually contain N-terminal peptides referred to as “transit peptides,” which are recognized by the translocation machinery. These transit peptides are predictable *in silico* (Table 10.1), and their detection can help the subcellular localization of the proteins carrying them. Based on the prediction models, the chloroplast proteome can be estimated at roughly 4500 different proteins. However, some plastid proteins lack a transit peptide [42] and many nuclear-encoded plastid proteins remain to be discovered. For example, proteomics analysis of the *Arabidopsis* chloroplast identified 690 different proteins [16]. Most of these proteins have been assigned to known protein complexes and metabolic pathways, but more than 30% of the proteins have unknown functions, and many are not predicted to localize to the chloroplast. To generate a complete catalogue of the chloroplast proteome, further fractionation into smaller subproteomes is required. In recent years, a number of comprehensive proteomics studies have focused on the chloroplast envelope [15, 43, 44], the thylakoid lumen [14, 45, 46], and thylakoid membrane [21, 47]. However, it is unlikely that any of these studies have been exhaustive, and low-abundance proteins remain to be detected. To date, around 1300 proteins have been characterized using MS-based studies (Figure 10.2, Table 10.1), and this observation must encourage further analysis of subproteomes and development of methods allowing identification of low abundance proteins.

Mitochondria, another double-membrane organelle, play a central role in eukaryotic cells by providing ATP through the oxidative phosphorylation process, but are also involved in many other cellular functions. When compared to their animal counterparts, plant mitochondria perform additional functions that cover a large panel of biochemical reactions and physiological processes (photorespiration step, implication in biosynthesis of essential cofactors such as vitamin C, biotin, folate, and lipoic acid). The *Arabidopsis* mitochondrial proteome was first analyzed by 2-DGE, and reference

maps were established [17, 18]. This led to the identification of 52 proteins [18] and 91 proteins [17] corresponding to the majority of the well-known relatively abundant enzymes of mitochondrial metabolism (Table 10.1, Figure 10.2). Methods combining extraction of hydrophobic proteins by organic solvents and salt combined with (1D) SDS-PAGE and MS/MS have allowed identification of intrinsic membrane proteins including carrier proteins and subunits of mitochondrial protein import machinery [19, 48]. A more detailed investigation has allowed characterization of a set of 419 proteins, nearly 80% of which were identified proteins that could be assigned to functional categories. A significant number of these low-abundance proteins are involved in DNA synthesis, transcription regulation, or protein complex assembly [49, 50]. Taken together, these proteomic works have allowed the identification of 547 nonredundant mitochondrial proteins, a number that is far below the 3200 mitochondrial proteins predicted by TargetP (Table 10.1).

Organelles with Single Membranes: Peroxisomes and Vacuoles. The function of the peroxisome found in all eukaryotes is to rid the cell of toxic substances. Before Fukao et al.'s study in 2002 [15], only a few peroxisomal proteins had been identified and characterized. In this study, the proteomic analysis of peroxisomes purified from greening cotyledons was undertaken, leading to the identification of 42 proteins by MALDI-TOF-MS. For the first time, protein kinases and protein phosphatases were identified as peroxisomal proteins, suggesting that protein phosphorylation could occur as a regulatory mechanism in leaf peroxisomes.

Plant cell vacuoles are multifunctional organelles that play a key role in plant physiology. Vacuoles can occupy up to 90% of the cellular volume in mature cells and are involved in the storage of a plethora of metabolites essential for plant metabolism, including water, inorganic anions and cations, organic and amino acids, sugars, and also proteins and insoluble compounds. The vacuole is also used for storage of xenobiotics or the by-products of metabolism which could become harmful if they accumulated in other cell compartments. To gain insight into vacuole functions, *Arabidopsis* "vacuomes" (vacuolar proteome; see also Chapter 27) were established by several groups, and 1038 different proteins were identified (Figure 10.2) [28, 52–54]. In one case, confirmation of vacuolar localization of GFP-fusion proteins was performed [28]. Proteomics study of the vacuole is one of the more complicated cases of plant organelles because of its role in the autophagic process. Because this organelle plays an essential role in bulk protein turnover and in the autophagic engulfment of cytoplasm and organelles (especially during senescence, starvation, and growth processes), it is difficult to discriminate the genuine organelle resident proteins from the contaminating ones. In order to circumvent such difficulties, Dunkley and co-workers have conceived an elegant method called LOPIT for localization of organelle proteins by isotope tagging based on isotope tags for relative and absolute quantification combined with 2D LC-MS/MS. This technique has allowed the discrimination between ER and Golgi apparatus-localized proteins [55]. More recently they applied this technique to organelles distributed along a density gradient, and 527 of 689 identified proteins were assigned to the ER, Golgi apparatus, vacuolar membrane, plasmalemma, mitochondria, or plastids. Location of some of these proteins was confirmed by GFP protein construct [56].

Most of the proteomics data obtained on *Arabidopsis* organelles are available on-line in the subcellular proteomic database (SUBA) (Table 10.1). This database also contains results from microscopy analysis of GFP fusion constructs, several independent bioinformatics prediction programs for protein subcellular localizations, and several other useful tools.

Differential Expression Proteomics

With the emergence of systems biology, it is increasingly clear that proteomics must undergo a paradigm shift from the descriptive area to differential expression proteomics. Several quantitative proteomics studies have been done to characterize the proteome changes when a tissue or a cell has been subjected to an exogenous stimulus or to analyze modifications of the protein profile along different developmental stages.

Proteomics Studies of Developmental and Cellular Processes. One of the most extensive proteomics works concerning the study of plant development was carried out on seed to analyze the germination process. These studies were performed to identify specific marker proteins for the different germination stages in order to better characterize the biochemical and molecular events underlying this developmental process [8]. 2-DGE analyses showed that the intensity of 1251 of 1272 proteins analyzed remained constant throughout the germination process (i.e., up to 2 days of imbibition) and supported the idea that dry seeds are essentially ready to germinate. The proteomic work of Gallardo et al. [57] highlighted enzyme markers that play a key role in the metabolic control of seed maturation and germination. The same group has also examined the role played by the gibberellin (GA) hormones in germination. Using a specific inhibitor of GAs biosynthesis and the GA-deficient *gal* mutant, they have shown that GAs do not participate in many processes involved in germination *sensu stricto*. Proteomics has also been used to analyze programmed cell death (PCD) in plants [58]. In contrast to animals where this cellular suicide is well-studied, little is known in plants. The work of Swidzinski et al. [58] has allowed the identification of a number of proteins potentially involved in plant PCD. Functional analyses are necessary to determine the role played by these candidate proteins in this process.

Proteomics as a Tool to Characterize Mutants and Genotypes. Several studies have demonstrated that genetic relationships between species and genera can be established based on the comparison of protein patterns [reviewed in references 50 and 60]. One of the first characterizations of *Arabidopsis* mutants using 2D gels was done by Santoni et al. [61]. 2-DGE analyses of proteins extracted from mutants and wild-type plants cultivated in the presence of various hormones, combined with statistical data treatment, allowed the estimation of biochemical distances between different genotypes. These distances were correlated with phenotypical and physiological analyses of the mutants, showing that this approach can be helpful in characterizing developmental mutants. 2-DGE has also been used as a tool to investigate global proteome variation among *Arabidopsis* ecotypes in order to estimate the biodiversity of these plant species [62–64]. Other proteomics studies have characterized

Arabidopsis mutant genotypes [57, 65]. Sorin et al. [65] studied mutants affected in their ability to develop adventitious roots and identified 11 proteins as potential markers of adventitious root development. In Wang et al. [66], flower proteome comparison between *Arabidopsis* wild-type and an *ask1* mutant was performed in order to investigate proteins that are affected by the ASK1-mediated proteolysis pathway. The analysis showed that proteins involved in photomorphogenesis, circadian oscillation, post-translational, stress response, cell expansion, or elongation processes were affected in the *ask1* mutant.

Environmental Proteomics. Plants are sessile organisms unable to escape environmental pressures. As a result, they have evolved an array of mechanisms and strategies to tolerate environmental conditions. How plants respond to stresses such as drought, high or low temperatures, high light, nutrition deficiencies, pathogen attacks, heavy metals, or other pollutants is crucial to improve crop productivity or plant's capacity to resist these aggressions. A number of proteomics studies have been conducted to analyze the impact of such environmental changes at the protein level. These proteomics investigations are generally based on a 2-DGE analysis of proteins extracted from treated or untreated plants, followed by gel image analyses to quantify and compare protein levels. Then, proteins up- and down-regulated in response to the treatment are identified by MS (MALDI and/or ESI-MS/MS). Table 10.2 summarizes the differential expression studies concerning abiotic and biotic stresses on *Arabidopsis*. Among the different stresses studied, there are: effect of SA on seed germination [67]; high light response of the thylakoid proteome in *Arabidopsis* wild-type and the ascorbate-deficient mutant *vtc2-2* [68]; cold stress response [69]; response to different nutrient culture conditions [13]; response to salt and osmotic stress [70]; response to bacterial pathogen [71] and response to cadmium (Cd) and cesium (Ce) [72, 73]. Some studies were carried out by analyzing proteome fluctuation at the tissue level [69, 74], at the cell level [13, 72, 73] or at the organelle or suborganelle level [38, 68, 75]. In some cases, the dynamics of the proteomes in response to a stress were explored with complementary approaches including metabolite profiling analyses performed by LC-MS/MS [13] or by nuclear magnetic resonance (NMR) spectroscopy [72].

Limitations of quantitative proteomics using 2-DGE become problematic with respect to certain classes of proteins such as integral membrane proteins or when the dynamic range of proteins is too high to be displayed on a gel. Promising gel-free technology using mass tagging approaches has been developed to overcome these disadvantages. These approaches are based on stable isotope labeling technology (SILAC) by both metabolic SILAC [76] and chemical methods including: formaldehyde [77], isotope-coded protein label (ICAT) [78], and ^{18}O -labeling [79].

MS: An Essential Tool for the Studies of PTM

Most proteins do not exist as individual functional entities but are involved in regulatory or metabolic networks. These proteins must interact with each other and PTM are one of the ways that proteins transmit signals. These modifications, such as phosphorylation, acetylation, myristoylation, glycosylation, ubiquitylation, or proteolytic

TABLE 10.2. Differential Expression Proteomics Studies Carried Out on *Arabidopsis* Following Biotic or Abiotic Treatments

Treatment, Time of Treatment	Arabidopsis Material	Levels of the Analysis	Proteomics Technologies	Other Analyses	References
Low temperature, 4°C, 6 h	3-week-old plants (Col-O), leaves	Nucleus	2-DE + MALDI	Northern blot analysis	38
Low temperature, 6°C and 10°C, 1 week	5-week-old plants (Col-O), all plant	Leaf	2-D DIGE + MALDI + ESI-MS/MS	None	69
Low temperature, 5°C Cold shock, 1 day, Short-term acclimation, 10 days; Long-term acclimation, 40 days	30-day-old plants (Col-O), old and new leaves	Chloroplast stroma and lumen	2-D DIGE + MALDI	None	128
Oxidative stress					
H ₂ O ₂ (88 mM), 16 h; Menadione (400 µM), 16 h; Antimycin A (25 µM), 16 h	Cell culture (Lansberg erecta)	Mitochondria	2-DE + Q-TOF MS	Respiratory measurements, Immunodetection of marker proteins, protease activity measurements	75
High light					
1000 µmol photons/m ² ; s 1, 3, and 5 days	40-day-old wild-type plants (Col-O) and ascorbate-deficient mutant Vtc2-2	Peripheral and luminal thylakoid	2-DE + ESI-MS/MS	Ascorbate and anthocyanin contents, Immunodetection	68

Red, far-red- and blue light				
Red (660 nm), 15 $\mu\text{mol m}^{-2} \text{s}^{-1}$, 6 $\mu\text{mol/m}^2 \text{s}^{-1}$ h; Far-red, (730 nm), 15 $\mu\text{mol/m}^2 \text{s}^{-1}$, 6 h; Blue (450 nm), 10 $\mu\text{mol/m}^2 \text{s}^{-1}$, 6 h	5-day-old wild-type seedlings (Col-O) and photoreceptor mutants: phyA-211, phyB-9 and phyAphyB	Seedling	2-DE + MALDI	Clustering analysis 129
Clinostat rotation (Gravity)				
Horizontal (H, simulated weightlessness) and vertical (V, clinostat control) Clinostat rotation and stationary control (S) growth condition	Callus cells (Lansberg erecta)	Cellular	2-DE + LC-IT-MS	Glucose, fructose, and starch measurements 130
K⁺ deficiency				
Normal (2 mM K ⁺) & K ⁺ -deficient medium (10 μM K ⁺), 3 h and 7 days	7-day-old seedlings (Col-O) (hydroponic culture)	Seedling	2-DE + MALDI	None 131
Nutrition				
Comparison of cells cultured in 2 media: Gamborg (GB5) & Murashige and Skoog (MS) medium	Cell culture (Colombia)	Cellular	2-DE + MALDI + ESI-MS/MS	None 13

(continued)

TABLE 10.2. (Continued)

Treatment, Time of Treatment	Arabidopsis Material	Levels of the Analysis	Proteomics Technologies	Other Analyses	References
Cadmium					
Cells grown in presence of CdCl ₂ (0, 0.5, 2, 5, 20, 50, and 200 μM) and in 2 nutrition conditions (GB5 & MS), 24 h	Cell culture (Colombia)	Cellular	2-DE + MALDI + ESL-MS/MS	RNA blot & RT PCR, Metabolite profiling	73
Cadmium					
Plants grown in presence of 10 μM Cd ²⁺ , 24 h	6-week-old plants (Col-O), roots (hydroponic culture)	Root	2-DE + MALDI	Western blot, glutathione affinity chromatography	74
Cesium					
Cells grown in presence of potassium (20 mM K; 0 mM Cs) versus cells treated with 1 mM cesium in presence or in absence of potassium, 24 h	Cell culture (Colombia)	Cellular	2-DE + MALDI	Chlorophyll content measurement, <i>in vivo</i> monitoring of Cs by NMR, synchrotron-based X-Ray fluorescence microscopy, atomic emission spectrometry, metabolite profiling	72
Pathogen					
<i>Pseudomonas syringae</i> pv. Tomato DC3000; DC3000 (avrRpm1) & DC3000(hrpA); 1.5, 3, 4, and 6 h after inoculation	5-week-old plants	Leaf	2-DE + Q-TOF MS	GeneChip	71

Pathogen elicitor, Fusarium elicitor (400 µg/mL), 24 h	Cell culture	Cellular	2-D DIGE + MALDI	Apoplast alkalization and H ₂ O ₂ accumulation, real-time PCR	132
Ethylene Germination in presence or not of aminooxyacetic acid (AOA, 250 µM, ACC synthase inhibitor) followed by treatment or not with ACC (100 µM)	4-day-old seedlings (Columbia)	Root, specific target enzyme families	2-DE + MALDI	Root hair formation, RNA blot	133
Salicylic acid (SA) Cells grown in presence of 0, 0.01, 0.1, and 1 mM SA	Cell culture	Cellular and specific target enzyme families	2-DE + LC-MS/MS	Western-blot, Glutathione affinity chromatography, Real-time PCR	124
Salicylic acid Seed germination in presence of SA, 500 µM, 1 day	Seeds after 1 day of imbibition	Seed	2-DE + MALDI	Influence of SA, sulfo-SA, and acetyl-SA on germination in optimal conditions or under salt stress, Detection of oxidized proteins and metabolic labeling	67

processing, occur after protein synthesis. They modulate enzymatic activity, binding ability, subcellular localization, how long proteins remain active, and so on. Protein modification by reversible phosphorylation is the most extensively studied PTM, and it is involved in almost all signaling pathways in plants [80]. The regulation of such an extensive and dynamic phosphorylation status mobilizes more than 5% of the *Arabidopsis* genome (approximately 1100 protein kinases and between 100 and 200 protein phosphatases) [81, 82], allowing a large degree of combinatorial regulation at the post-translational level (see PlantsP, Table 10.1). Phosphorylated proteins have been characterized using classical tools such as incorporation of radioactive orthophosphate by cells submitted to bacterial or fungal elicitors, followed by 2-DGE separation and MS identification of labeled proteins [83, 84]. Phosphoaffinity column chromatography combined with the use of phosphatase treatment and phosphospecific staining has allowed the characterization of a set of phosphorylated late embryogenesis abundant (LEA) proteins in mature *Arabidopsis* seed [85]. To decipher the phosphoproteome, sophisticated approaches using IMAC charged with transition metals (Fe^{3+} or Ga^{3+}) were developed in order to selectively purify phosphopeptides from complex mixtures. Using this approach, the evolution of the phosphorylation status of different photosynthetic proteins under several physiological conditions were analyzed [86–88]. Phosphorylation sites of proteins involved in RNA metabolism were also determined [87]. The first successful large-scale IMAC study published led to the identification of more than 300 phosphorylation sites in *Arabidopsis* PM proteins [89, 90] (Table 10.2). A potential downfall of IMAC technology is that it may also bind peptides containing acidic residues, and methylation of these acidic residues can help to improve the specific binding of phosphopeptides [88]. Recently, Wolschin and Weckwerth [91] have developed a promising procedure for the enrichment of phosphorylated proteins and peptides from complex mixtures based on the specific interaction of the phosphate group with metal oxide/hydroxide affinity chromatography (MOAC). A growing number of observations show that protein phosphorylation occurs at different sites in a single protein [92–94]. Moreover, the modifications of multiple sites constitute an additional layer of molecular information beyond the amino acid sequence. These modifications are coordinated, reversible, and kinetically regulated, and they work in tandem to control innumerable cellular functions. Specific methods are required to analyze the dynamics of these events in a cellular context and have been developed to quantify them. To this end, Hegeman et al. [95] have developed a method of determining phosphorylation stoichiometries by using methyl-esterification with an isotopic labeling reagent (methyl alcohol- d_4). Using the increasing technological capacities of MS technology, Glinski and Weckwerth [93] have developed a highly selective LC–MS/MS-based method using high-resolution multiple reaction monitoring on a triple quadrupole mass spectrometer, to eliminate the need for SIL. The strategy combines enrichment of phosphoproteins using the MOAC technique and selective ion trap MS, providing high confidence in the identification of phosphoproteins and their corresponding phosphorylation sites.

With the objective of revealing novel post-translational regulations of plant proteins, several groups have recently engaged a thorough analysis of the modification status in various *Arabidopsis* tissues, cells or organelles. Santoni et al., [96] have

shown that the PIP aquaporins of the PM are methylated. PM proteins are also modified by GPI. The GPI-anchored proteins (GAPs) are targeted to the plant cell surface and are likely to be involved in extracellular matrix remodeling, cell adhesion, differentiation, and host–pathogen interactions. Proteomics works have allowed the identification of several GAPs [27, 97, 98]. PTM by *N*-myristoylation has also been studied and the putative *Arabidopsis* “*N*-myristoylome” consists of 437 proteins, accounting for 1.7% of the complete proteome [99]. A growing body of evidence suggests that protein *S*-nitrosylation, the covalent attachment of nitric oxide (NO) to the thiol group of Cys residues, plays a regulatory role in animals but also in plants. Lindermayr et al. [100, 101] have identified more than 100 proteins as *S*-nitrosylation targets by the *in vitro* biotin switch method suggesting that a number of activities are regulated by *S*-nitrosylation. Reversible modification of Cys residues by *S*-glutathionylation was also recently studied in *Arabidopsis* and could play a role in redox regulation of protein activities [102].

From Functional Proteomics to Systems Biology: Use of Targeted Proteomic Approaches

Elucidation of the molecular mechanisms underlying cell functions requires a global knowledge of the expressed proteins at any given time and in various conditions during the life of the cell or under any cellular conditions. Global knowledge encompasses splice variant products, the level of their expression, their subcellular localizations, their PTMs, and the nature of their interacting partners. Because of the wide dynamic range of intracellular protein concentrations, protein enrichment strategies are necessary when proteomics studies are to be applied to the analysis of a set of proteins of interest—for example, members of multiprotein complexes. Before their identification by MS, these protein components have to be purified and several methods have been utilized to enrich them. When these complexes are present in an organelle, it is simplest to enrich the subcellular fraction of interest. Such approaches have been used to characterize the membrane protein components of PSI and II [103] and of mitochondrial respiratory complexes and supracomplexes [104–106]. In the latter studies, BN gel (either combined or not combined with classical 2-DGE) was performed in complement; in the first study, however, multidimensional analyses by HPLC–ESI–MS/MS were used. Soluble HMW complexes may be purified by the use of high-speed centrifugation followed by sucrose gradient density; such a procedure was used for the characterization of cytosolic ribosomes [107]. Polyethylene glycol (PEG) precipitation followed by purification steps on chromatographic columns was another alternative to analyze components of proteolytic complexes such as the *Arabidopsis* 26S proteasome [108] and tripeptidyl peptidase II [109]. Co-immunoprecipitation, another way to analyze protein interactions, was used to characterize the composition of polyribosomal complexes [110], somatic embryogenesis receptor-like kinase (RLK) 1 protein complex [111], phytochrome-interacting protein candidates [112], and chloroplast putative nitrogen sensor (PII) [113]. Recently, affinity chromatography has allowed the identification of putative PII-binding proteins [114]. Similar chromatography has also been used to characterize *in vitro* tubulin-binding proteins [115]. Alternatively, affinity

capture/enrichment techniques (TAP strategy) were used to examine supramolecular complexes or to analyze protein interactions [116–119]. As proof of concept, all these protein interactions must be investigated by other approaches such as immunofluorescence [119] or *in vivo* FRET [111, 120, 121].

Efforts have led to the development of tagged chemical probes that selectively react with families of active enzymes based on shared mechanistic features. This activity-based proteomic approach permits the comparison and possible quantitation of multiple protein activities simultaneously in complex proteomes. This technology was used to classify the 500 predicted *Arabidopsis* proteases (MEROPS, Table 10.1), based on their catalytic mechanisms, into Ser, Cys, Asp, and metalloproteases [122]. Glutathione-affinity chromatography was used to identify *Arabidopsis* GSTs and to analyze their expression in response to different stresses [123, 124]. A similar approach based on IMAC strategies was used to screen copper-interacting proteins [125]. Chemical strategies introducing small-molecule probes to profile dynamics of protein activity in complex proteomes have proven particularly valuable [100, 102, 126, 127]. For example, monobromobimane as a fluorescent probe combined with 2-DGE was used to identify proteins targeted by thioredoxin [127]. Another approach using iodoacetamide-based fluorescence tagging of proteins in conjunction with MS analysis led to the identification of several proteins that might be potential targets of hydrogen peroxide (H_2O_2) [126]. Experimentally and *in silico* predicted protein interactions are currently available on-line in databases such as BOND, IntAct, MINT, or DIP (Table 10.1).

10.4 CONCLUSIONS

Over the last five years, the number of publications on *Arabidopsis* proteomics has exploded. These publications cover a large range of proteomic aspects, from the straightforward proteome inventories to various features of functional proteomics. Nevertheless, the number of proteins characterized by proteomic approaches remains relatively low (<3500 proteins) when compared to the 26,000 genes present in the *Arabidopsis* genome. This must encourage the development of methods to analyze both (a) low-abundance proteins (especially membrane proteins) present in nucleus, plastid, mitochondrion, vacuole, cytosol, and other subcellular compartments and (b) the organization of protein complexes in plant cells in order to build a “cellular map” of the plant cell proteome. This “image” of the cellular proteome will be a fantastic tool to analyze the influence of a perturbation on the dynamics of the cellular and subcellular proteomes.

10.5 FIVE-YEAR VIEWPOINT

The huge quantity of information that can be obtained from high-throughput proteomics analysis needs to be further exploited. For this, development of specific biochemical tools to gain insight into the proteome and define macromolecular complex interactions at any given time and in various conditions during the life of the cell

or under any cellular conditions is required. Because of the wide dynamic range of intracellular protein concentrations, enrichment of low-abundance proteins is one of the issues that proteomics approaches must focus on in the future. However, the size and diversity of the data being generated is impossible to organize without computational assistance, and automatic analysis of the mass spectrometric data is required. Computational assistance is, however, still at an early stage of development. Effort must be put on database infrastructure, data storage and recovery, and interface design to generate model hypotheses that can be tested experimentally. Therefore, the development of so-called “systems biology” (see Chapter 46) combining and integrating experimental data from diverse sources, such as proteome, interactome, metabolome, and transcriptome, is the key issue and the challenge for the near future.

ACKNOWLEDGMENTS

We are grateful to Dr. Maighread Gallagher-Gambarelli for critical reading of the manuscript, and we thank Jacqueline Martin-Laffon for assistance in the bibliography research procedure.

REFERENCES

1. *Arabidopsis* Genome Initiative. 2000. *Nature* **408**:796–815.
2. Unseld M, et al. 1997. *Nat Genet* **15**:57–61.
3. Sato S, et al. 1999. *DNA Res* **6**:283–290.
4. Frishman D, et al. 2001. *Bioinformatics* **17**:44–57.
5. Haas BJ, et al. 2003. *Nucleic Acids Res* **31**:5654–5666.
6. Kamo M, et al. 1995. *Electrophoresis* **16**:423–430.
7. Giavalisco P, et al. 2005. *Proteomics* **5**:1902–1913.
8. Gallardo K, et al. 2001. *Plant Physiol* **126**:835–848.
9. Fu Q, et al. 2005. *J Biochem Mol Biol* **38**:650–660.
10. Noir S, et al. 2005. *Biochem Biophys Res Commun* **337**:1257–1266.
11. Holmes-Davis R, et al. 2005. *Proteomics* **5**:4864–4884.
12. Wienkoop S, et al. 2004. *Phytochemistry* **65**:1641–1649.
13. Sarry JE, et al. 2006. *Electrophoresis* **27**:495–507.
14. Peltier JB, et al. 2002. *Plant Cell* **14**:211–236.
15. Ferro M, et al. 2003. *Mol Cell Proteomics* **2**:325–345.
16. Kleffmann T, et al. 2004. *Current Biol* **14**:354–362.
17. Millar AH, et al. 2001. *Plant Physiol* **127**:1711–1727.
18. Kruff V, et al. 2001. *Plant Physiol* **127**:1694–1710.
19. Brugiere S, et al. 2004. *Phytochemistry* **65**:1693–1707.
20. Santoni V, et al. 2000. *Electrophoresis* **21**:3329–3344.
21. Friso G, et al. 2004. *Plant Cell* **16**:478–499.
22. Santoni V, et al. 2003. *Biochem J* **373**:289–296.

23. Luche S, Santoni V, and Rabilloud T. 2003. *Proteomics* **3**:249–253.
24. Prime TA, et al. 2000. *Electrophoresis* **21**:3488–3499.
25. Marmagne A, et al. 2004. *Mol Cell Proteomics* **3**:675–691.
26. Sazuka T, et al. 2004. *DNA Res* **11**:101–113.
27. Borner GH, et al. 2003. *Plant Physiol* **132**:568–577.
28. Jaquinod M, et al. 2007. *Mol Cell Proteomics* **6**:394–412.
29. Schwacke R, Flugge UI, and Kunze R. 2004. *Plant Physiol Biochem* **42**:1023–1034.
30. Jamet E, et al. 2006. *Trends Plant Sci* **11**:33–39.
31. Charmont S, et al. 2005. *Phytochemistry* **66**:453–461.
32. Kwon HK, Yokoyama R, and Nishitani K. 2005. *Plant Cell Physiol* **46**:843–857.
33. Chivasa S, et al. 2002. *Electrophoresis* **23**:1754–1765.
34. Borderies G, et al. 2003. *Electrophoresis* **24**:3421–3432.
35. Feiz L, et al. 2006. *Plant Methods* **2**:10.
36. Boudart G, et al. 2005. *Proteomics* **5**:212–221.
37. Bayer EM, et al. 2006. *Proteomics* **6**:301–311.
38. Bae MS, et al. 2003. *Plant J* **36**:652–663.
39. Pendle AF, et al. 2005. *Mol Biol Cell* **16**:260–269.
40. Brown JW, et al. 2005. *Nucleic Acids Res* **33**:D633–D636.
41. Nair R and Rost B. 2005. *J Mol Biol* **348**:85–100.
42. Miras S, et al. 2002. *J Biol Chem* **277**:47770–47778.
43. Eichacker LA, et al. 2004. *J Biol Chem* **279**:50915–50922.
44. Froehlich JE, et al. 2003. *J Proteome Res* **2**:413–425.
45. Kieselbach T, et al. 2000. *FEBS Lett* **480**:271–276.
46. Schubert M, et al. 2002. *J Biol Chem* **277**:8354–8365.
47. Peltier JB, et al. 2004. *J Biol Chem* **279**:49367–49383.
48. Millar AH and Heazlewood JL. 2003. *Plant Physiol* **131**:443–453.
49. Heazlewood JL, et al. 2004. *Plant Cell* **16**:241–256.
50. Heazlewood JL and Millar AH. 2005. *Nucleic Acids Res* **33**:D605–D610.
51. Fukao Y, Hayashi M, and Nishimura M. 2002. *Plant Cell Physiol* **43**:689–696.
52. Szponarski W, et al. 2004. *Proteomics* **4**:397–406.
53. Shimaoka T, et al. 2004. *Plant Cell Physiol* **45**:672–683.
54. Carter C, et al. 2004. *Plant Cell* **16**:3285–3303.
55. Dunkley TP, et al. 2004. *Mol Cell Proteomics* **3**:1128–1134.
56. Dunkley TPJ, et al. 2006. *Proc Natl Acad Sci USA* **103**:6518–6523.
57. Gallardo K, et al. 2002. *Plant Physiol* **129**:823–837.
58. Swidzinski JA, Leaver CJ, and Sweetlove LJ. 2004. *Phytochemistry* **65**:1829–1838.
59. Thielllement H, Zivy M, and Plomion C. 2002. *J Chromatogr B* **782**:137–149.
60. Agrawal GK, et al. 2005. *J Chromatogr B* **815**:137–145.
61. Santoni V, et al. 1997. *Planta* **202**:62–69.
62. Chevalier F, et al. 2004. *Proteomics* **4**:1372–1381.
63. Ruebelt MC, et al. 2006. *J Agric Food Chem* **54**:2162–2168.
64. Schlesier B, et al. 2004. *Phytochemistry* **65**:1565–1574.

65. Sorin C, et al. 2006. *Plant Physiol* **140**:349–364.
66. Wang XF, et al. 2006. *Cell Res* **16**:489–498.
67. Rajjou L, et al. 2006. *Plant Physiol* **141**:910–923.
68. Giacomelli L, Rudella A, and van Wijk KJ. 2006. *Plant Physiol* **141**:685–701.
69. Amme S, et al. 2006. *J Exp Bot* **57**:1537–1546.
70. Ndimba BK, et al. 2005. *Proteomics* **5**:4185–4196.
71. Jones AM, et al. 2004. *Phytochemistry* **65**:1805–1816.
72. Le Lay P, et al. 2006. *Biochimie* **88**:1533–1547.
73. Sarry JE, et al. 2006. *Proteomics* **6**:2180–2198.
74. Roth U, von Roepenack-Lahaye E, and Clemens S. 2006. *J Exp Bot* **57**:4003–4013.
75. Sweetlove LJ, et al. 2002. *Plant J* **32**:891–904.
76. Gruhler A, et al. 2005. *Mol Cell Proteomics* **4**:1697–1709.
77. Ytterberg AJ, Peltier JB, and van Wijk KJ. 2006. *Plant Physiol* **140**:984–997.
78. Dunkley TP, et al. 2004. *Biochem Soc Trans* **32**:520–523.
79. Nelson CJ, et al. 2006. *Mol Cell Proteomics* **5**:1382–1395.
80. Peck SC. 2006. *J Exp Bot* **57**:1523–1527.
81. Luan S. 2003. *Annu Rev Plant Biol* **54**:63–92.
82. Kerk D, et al. 2002. *Plant Physiol* **129**:908–925.
83. Peck SC, et al. 2001. *Plant Cell* **13**:1467–1475.
84. Nuhse TS, Boller T, and Peck SC. 2003. *J Biol Chem* **278**:45248–45254.
85. Irar S, et al. 2006. *Proteomics* **6**(Suppl 1):S175–S185.
86. Vener AV, et al. 2001. *J Biol Chem* **276**:6959–6966.
87. de la Fuente van Bentem S, et al. 2006. *Nucleic Acids Res* **34**:3267–3278.
88. Hansson M and Vener AV. 2003. *Mol Cell Proteomics* **2**:550–559.
89. Nuhse TS, et al. 2004. *Plant Cell* **16**:2394–2405.
90. Nuhse TS, et al. 2003. *Mol Cell Proteomics* **2**:1234–1243.
91. Wolschin F, Wienkoop S, and Weckwerth W. 2005. *Proteomics* **5**:4389–4397.
92. Furihata T, et al. 2006. *Proc Natl Acad Sci USA* **103**:1988–1993.
93. Glinski M and Weckwerth W. 2005. *Mol Cell Proteomics* **4**:1614–1625.
94. Hong-Hermesdorf A, et al. 2006. *FEBS Lett* **580**:932–939.
95. Hegeman AD, et al. 2004. *J Am Soc Mass Spectrom* **15**:647–653.
96. Santoni V, et al. 2006. *Biochem J* **400**:189–197.
97. Elortza F, et al. 2003. *Mol Cell Proteomics* **2**:1261–1270.
98. Elortza F, et al. 2006. *J Proteome Res* **5**:935–943.
99. Boisson B, Giglione C, and Meinel T. 2003. *J Biol Chem* **278**:43418–43429.
100. Lindermayr C, Saalbach G, and Durner J. 2005. *Plant Physiol* **137**:921–930.
101. Lindermayr C, et al. 2006. *J Biol Chem* **281**:4285–4291.
102. Dixon DP, et al. 2005. *Plant Physiol* **138**:2233–2244.
103. Huber CG, et al. 2004. *Proteomics* **4**:3909–3920.
104. Werhahn W and Braun HP. 2002. *Electrophoresis* **23**:640–646.
105. Dudkina NV, et al. 2005. *Proc Natl Acad Sci USA* **102**:3225–3229.
106. Millar AH, et al. 2004. *Plant Mol Biol* **56**:77–90.

107. Chang IF, et al. 2005. *Plant Physiol* **137**:848–862.
108. Yang P, et al. 2004. *J Biol Chem* **279**:6401–6413.
109. Book AJ, et al. 2005. *Plant Physiol* **138**:1046–1057.
110. Zanetti ME, et al. 2005. *Plant Physiol* **138**:624–635.
111. Karlova R, et al. 2006. *Plant Cell* **18**:626–638.
112. Phee BK, et al. 2006. *Proteomics* **6**:3671–3680.
113. Smith CS, et al. 2002. *Protein Expr Purif* **25**:342–347.
114. Chen YM, et al. 2006. *J Biol Chem* **281**:5726–5733.
115. Chuong SD, et al. 2004. *Mol Cell Proteomics* **3**:970–983.
116. Witte CP, et al. 2004. *Plant Mol Biol* **55**:135–147.
117. Rubio V, et al. 2005. *Plant J* **41**:767–778.
118. Brown AP, et al. 2006. *J Exp Bot* **57**:1563–1571.
119. Park NI and Muench DG. 2006. *Planta* **255**:625–639.
120. Seidel T, Gollack D and Dietz KJ. 2005. *FEBS Lett* **579**:4374–4382.
121. Aker J, et al. 2006. *J Struct Biol* **156**:62–71.
122. van der Hoorn RAL, et al. 2004. *Plant Physiol* **135**:1170–1178.
123. Smith AP, et al. 2004. *J Biol Chem* **279**:26098–26104.
124. Sappl PG, et al. 2004. *Plant Mol Biol* **54**:205–219.
125. Kung CCS, et al. 2006. *Proteomics* **6**:2746–2758.
126. Hancock JT, et al. 2005. *Plant Physiol Biochem* **43**:828–835.
127. Yano H, et al. 2001. *Proc Natl Acad Sci USA* **98**:4794–4799.
128. Goulas E, et al. 2006. *Plant J* **47**:720–734.
129. Kim DS, et al. 2006. *Proteomics* **6**:3040–3049.
130. Wang H, et al. 2006. *J Exp Bot* **57**:827–835.
131. Kang JG, et al. 2004. *Proteomics* **4**:3549–3559.
132. Chivasa S, et al. 2006. *J Exp Bot* **57**:1553–1562.
133. Mang HG, et al. 2004. *Biochim Biophys Acta* **1676**:231–239.

RICE PROTEOME AT A GLANCE

Ganesh Kumar Agrawal and Randeep Rakwal

11.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

“Rice is life” is a particularly apt saying for the millions of people across Southeast Asia and around the world, who depend upon this crop for the majority of their daily food needs. Indeed, it has been estimated that rice provides more than one-fifth of the calories consumed worldwide by humans [1]. It is perhaps not-surprising therefore, that rice is one of the most well studied plants in all of human scientific research.

Farmers are perhaps the original rice researchers. They have amassed a great deal of knowledge since rice was first domesticated via the breeding of new varieties and the study of meteorological and environmental factors to improve productivity and avoid disease, and so on. However, in more recent times, rice has spawned its own field of academic research in universities, institutes, and agricultural companies, around the world. Perhaps the most prominent among these is the International Rice Research Institute (IRRI), established at Los Banos in the Philippines in 1960 by the Ford and Rockefeller Foundations, in cooperation with the Philippine government. One of the most tangible outputs of modern rice science is the rice genome project(s) initiated by various private companies (most notably Monsanto and Syngenta), institutes, and governments. The IRGSP, a consortium of publicly funded laboratories, was established

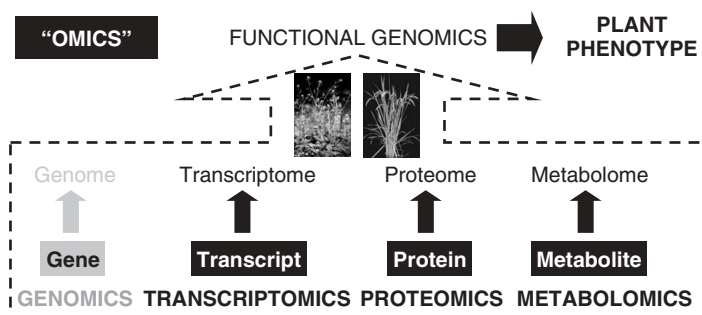


FIGURE 11.1. OMICS in plant biology.

in 1997 with the aim of obtaining a high-quality, map-based sequence of the rice genome using the Nipponbare cultivar of the *O. sativa* ssp. japonica rice strain [2].

Aside from genetic analysis (genomics), rice researchers also have a huge range of related techniques available to them, especially if they are involved in the study of gene transcripts (transcriptomics), proteins (proteomics), and metabolites (metabolomics). The genome forms the backdrop for advances in these techniques (Figure 11.1). The rice genome itself is almost fully annotated, but the most challenging task—namely, the assignment of the function of each gene in order to build up a greater understanding of the plant phenotype—remains unfinished. This is a new frontier in plant science and will require systematic, high-throughput approaches.

We believe that the collective information from integrative “omics” technologies will help to provide the answers to the most fundamental question in plant biology, namely, How do biological systems work and interact with their environment [3, 4]? We have therefore previously dubbed rice as “a cornerstone for cereal food crop proteomes” [3] (Figure 11.2). Similarly, Dr. Robin Buell of The Institute for Genomic Research (TIGR) has called rice the “Rosetta stone for crop genomics” and “a critically important crop,” while TIGR President Claire Fraser recently remarked that “much as the Human Genome Project has revolutionized biology, the rice genome promises to inspire new cereal crop research, this is a major step forward for agriculture.” It is our belief that the sequenced rice (*O. sativa* L.) genomes (ssp. Japonica and Indica) are indeed a 21st-century landmark in the history of plant biology and agriculture.

Proteomics is one of the most promising and exciting “omics” approaches. This technique has been applied to a variety of plant species with the aim of identifying and profiling the proteins expressed in plant tissues or organelles. This approach has been most fully applied in the dicot *A. thaliana* (L.) Heyhn, which was the first plant to have its genome fully decoded [5–9]. A simple PubMed database search using the keyword “plant and proteome” gives 798 hits (as of October 1st, 2007), revealing the interest and scope of proteomics in plant biology. A recent comprehensive review [10] updates us on the contribution of proteomics to plant biology, and the interested reader is referred to it for the same purpose. However, as the authors of that review point out, there has been a quantitative (in terms of publications), but not a qualitative, leap in

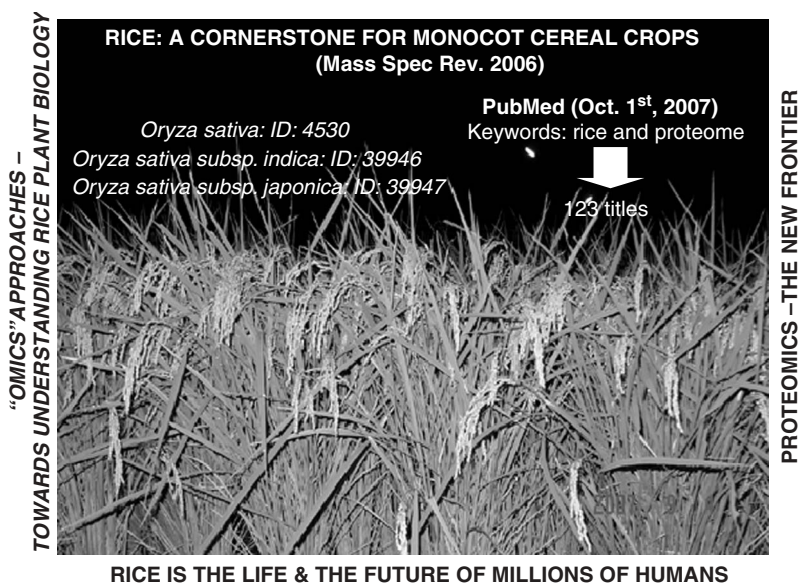


FIGURE 11.2. Rice, *Oryza sativa* L., is a plant model for functional genomics of cereal crops.

plant proteomics [10]. This is understandable, but we still have a long road ahead if we are to utilize existing technology to develop the “omics” sciences to their full potential.

Our own comprehensive and detailed reviews in rice have dealt with the early stages of rice proteomics (in 2003) [11], and in 2006 we reviewed the state of the science today [3, 4]. In the following sections we deal with the methodology and strategy, detailing what has been achieved in rice proteomics up to the present and discussing what has the potential to be achieved in the future.

11.2 METHODOLOGY AND STRATEGY

Technology and ideas are the engines of growth and thus the driving forces behind scientific advancements. This is true for all the “omics” sciences and particularly proteomics. Today, proteomics can boast of a host of technologies and equipments/instruments which make it a leading and exciting research field to be a part of. Proteomics began by using very simple gel-based approaches, such as 1-DGE and 2-DGE (though mostly 2-DGE), for which readers are referred to Chapter 2 for more details. If the gel-based approach is one pillar of proteomics, the other is undoubtedly MS, which is covered in great detail in Chapter 3. For readers specifically interested in the other technologies used in rice proteomics, there are two comprehensive reviews available [3, 4]. Moreover, several chapters (6, 16, 17, 24, and 36) in this book also provide further background details.

As we all know, proper protein extraction of the sample (cell, tissue, organ, or indeed whole plant) in question is essential before any meaningful proteome study can be developed. The original technique developed by O'Farrell in the 1975 [12] is suitable for protein extraction using LB from any raw material. However, many researchers have adapted, modified, and improved various extraction protocols specifically for rice. In our experience, the use of trichloroacetic acid/acetone extraction buffer (TCAEAB) along with LB-TT [4] is highly advantageous (especially for 2-DGE).

An optimized protein extraction protocol for rice leaves is schematically presented in Figure 11.3. It should be noted that the solubilized protein (in LB-TT) can be processed for both 1-DGE and 2-DGE (with or without use of the 2-D Clean Up Kit, GE Healthcare Bio-Sciences AB). The composition of the solutions used in Figure 11.3

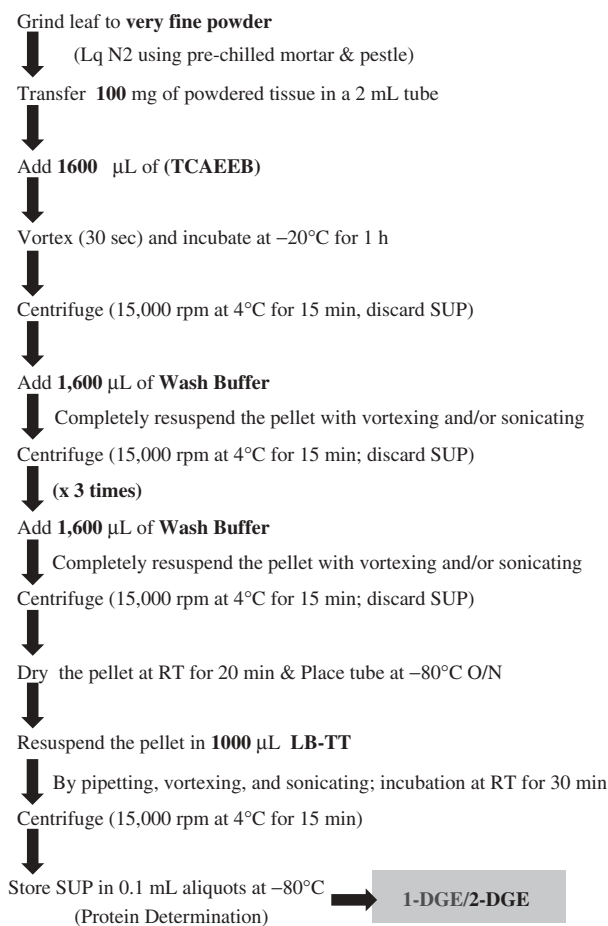


FIGURE 11.3. Standardized protein extraction protocol for rice leaf using TCAEAB protocol. The 2-D Clean-Up Kit from GE Healthcare kit may be used before 2-DGE.

TCA/acetone ext. buffer (TCAEB)		
10% (w/v) TCA		10 g
0.07% β-mercaptoethanol		70 μL
Acetone		final vol. 100 ml
Wash Buffer (WB)		
0.07% β-mercaptoethanol		70 μL
2 mM EDTA		400 μL
	(500 mM EDTA stock)	
EDTA-free Proteinase Inhibitor		1 tablet
Acetone		final vol. 100 mL
LB-TT(+ 50 mM DTT)		
Urea	7 M	42.0 g
Thiourea	2 M	15.2 g
CHAPS	4% (w/v)	4.0 g
1 MTris-HCl (pH8.0-8.3)	18 mM	1.8 mL
Trizma Base	14 mM	69.5 mg
EDTA-free Proteinase Inhibitor		2 tablets
Triton X-100®	0.20%	200 μL
DTT	50 mM	71.5 mg
Autoclaved MQ		final vol. 100 mL

FIGURE 11.4. Composition of the TCAEB solution, the wash buffer, and the LB-TT (resolubilization of pellet and rehydration/focusing solution) used for rice leaf protein extraction.

is given in Figure 11.4. Recent results also demonstrate that the TCAEB/LB-TT protocol can significantly improve the solubilization of proteins from not only rice tissues, but also from *Arabidopsis*, maize, cucumber, and wheat seed embryos [13]. Researchers also have a wide choice of buffers to use and adapt for a particular sample, as evident from many research publications. In this regard, we also recommend using the LB-TT (direct extraction) or the phenol buffer. The former gives excellent extraction efficiency while avoiding loss of proteins, while the latter is useful for removing contaminants (during IEF) such as nucleic acid, phenols, polysaccharides, and so on. Ultimately, however, a researcher can use the extraction protocol of their choice by modifying and adapting available protocols to extract the best quality and composition protein, free of contaminants.

The second critical step for developing a proteome reference map (via 2D gel analysis) is the focusing and separation of extracted protein in both the first and second dimensions. Although 2-DGE has inherent drawbacks, such as underrepresentation of highly acidic, basic, hydrophobic, and extremely high- and/or low-molecular-weight (HMW/LMW) proteins, it is the method of choice for most laboratories (its merits/demerits are further debated in Chapter 2). In rice, the hand-cast IEF system has long been the method of choice for many research groups mainly due to cost-effectiveness [3, 4]. We feel that a shift to pre-cast IPG technology and large-format gels is a necessity. Most laboratories and research groups focused on rice proteome research, especially those developing high-quality 2D reference gel maps, are already

using this technology [4]. Furthermore, we have standardized the IEF protocol for IPG strip formats (11–24 cm in length; GE Healthcare) using rice leaf proteins as an example (this is presented in Figure 11.5). This protocol can be further modified according to each specific sample and condition if required. With the use of a 24-cm IPG strip (pH 4–7), ~910 protein spots stained with silver nitrate (for a standardized staining protocol, see Table 11.1) were detected in rice leaves on pre-cast 12.5% DALT large-format PAG (Figure 11.6). The pre-cast IPG strips are therefore recommended not only for the better resolution of protein spots but also for greater reproducibility in generating high-quality 2D gel reference maps [4, 14].

Once a high-quality 2D gel has been obtained, the last two steps in the completion of a proteomics study are image analysis (see Chapters 2 and 8) and protein identification. For identifying the stained proteins spots, one can use either N-terminal amino

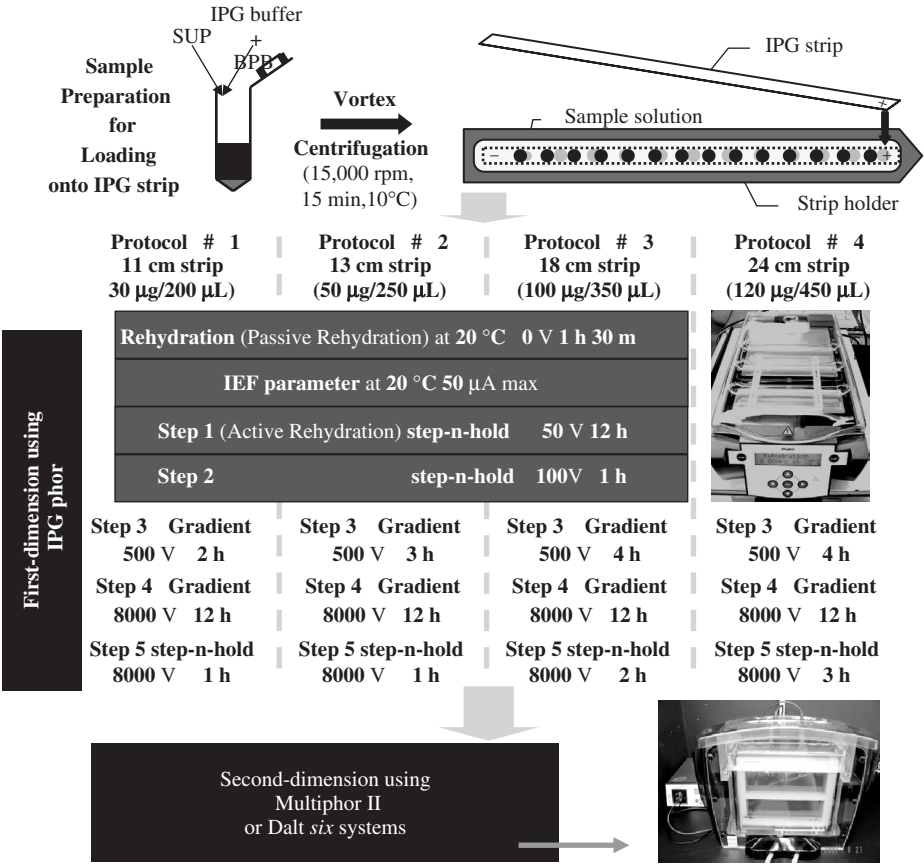


FIGURE 11.5. Optimized IEF parameters employed for first-dimension focusing of rice leaf proteins using the IPGphor equipment (GE Healthcare). Pre-cast IPG strips from GE Healthcare were used for the optimization of IEF conditions and examining the results via second dimension SDS-PAGE using pre-cast PAG.

TABLE 11.1. Optimized Silver Nitrate Staining* Protocol for Pre-Cast Gel (DAIT Gel 12.5; 255 × 196 × 1 mm)

Step	Solution	Volume (mL/gel)	Time
1. Fixation	40% (v/v) EtOH, 10% (v/v) acetic acid	200	120 min
2. Sensitizing	30% (v/v) EtOH, 0.125% glutaraldehyde, 0.2% sodium thiosulfate, 6.8% sodium acetate	250	120 min
3. Washing	Distilled water/MQ	300	5 × 15 min
4. Silver reaction	0.25% Silver nitrate, 0.0148% formaldehyde	250	120 min
5. Washing	Distilled water/MQ	300	4 × 1 min
6. Developing	2.5% Sodium carbonate, 0.0296% formaldehyde	250	4–6 min
7. Stopping	1.46% EDTA-Na ₂ · 2H ₂ O	250	60 min
8. Washing	Distilled water/MQ	300	2 × 15 min

*PlusOne Silver Staining Kit, Protein (Cat. No. 17-1150-01, GE Healthcare).

acid sequencing [15, 16] or MS analyses (Chapter 3). Although different research groups may have different protocols for MS analysis, essentially, in-gel digestion of excised protein spots from the 2D gel is usually performed using specific endopeptidases (e.g., trypsin). For rice proteins stained with silver nitrate (Figure 11.6), we have standardized an in-gel digestion protocol for the excised spots (Figure 11.7). Protein identification with MS is usually performed in a database-dependent/independent manner. Therefore MS, being high-throughput in nature, has almost replaced Edman sequencing as the method of choice for protein identification in proteomics. Nevertheless, Edman sequencing is still useful for the precise determination of N-terminals, and amino acid sequencing is still used in conjunction with MS for protein identification in the majority of proteomics projects. A strategy for rice proteomics that can be easily adapted for any project is presented in Figure 11.8.

As discussed in earlier reviews, bioinformatics and databases are an essential part of any proteomics project [3, 4]. A robust centralized and freely accessible rice resource database with wide ranging bioinformatics is provided by the TIGR. This is based on the public draft generated by the IRGSP (<http://www.tigr.org/tdb/e2k1/osa1/>) and includes the anchoring of the publicly available rice bacterial artificial chromosome clones to the genetic map. Recently, an update of the rice genome sequence with new data and improved annotation quality has become available as part of the TIGR Rice Genome Annotation project (<http://rice.tigr.org>). In the current annotation (release 4.0; January 12, 2006), 42,653 nontransposable element-related genes encoding 49,472 gene models as a result of the detection of alternative splicing have been identified. By incorporating multiple transcript and proteomic expression data sets, it has proved possible to annotate 24,799 genes (31,739 gene models), representing approximately 50% of the total gene models. For viewing all structural and functional annotation the readers are referred to the Rice Genome Browser, which currently supports 59 tracks. Enhanced data access is available through web

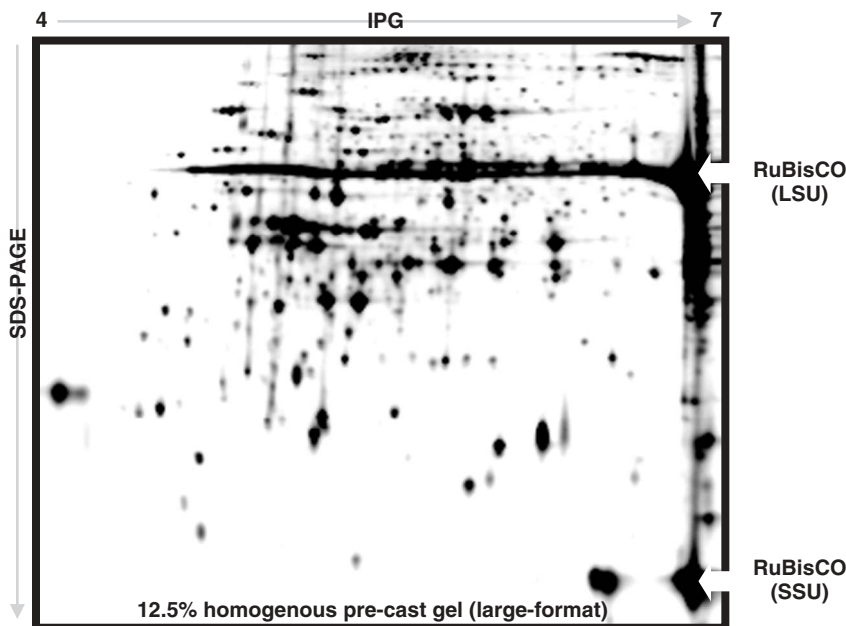


FIGURE 11.6. Representative high-resolution 2-D gel image of rice (cv. Nipponbare) leaf total soluble proteins. Protein (150 μ g) was separated on 24-cm pre-cast IPG strip gel (pH 4–7) and large format 12.5% PAG using IPGphor and Ettan DALT six systems from GE Healthcare Bio-Sciences AB. Separated protein spots on 2D gels were visualized with silver nitrate.

interfaces, FTP downloads, and a data extractor tool developed in order to support discrete dataset downloads [17]. In addition, the collaborative full-length cDNA project database is accessible through the knowledge-based *Oryza* Molecular biological encyclopedia website (KOME, <http://cdna01.dna.affrc.go.jp/cDNA/>) administered by the National Institute of Agrobiological Sciences (NIAS), Tsukuba, Japan.

11.3 EXPERIMENTAL RESULTS AND APPLICATIONS

Research Accomplishments Upto 2006

Although rice proteomics research began in 1993, it was only in the year 2000 that the field really began to gather momentum. In the last six years, numerous studies have been performed. These have involved mainly the developmental and environmental proteomics of rice and have been thoroughly reviewed recently [3, 4]. The information on these studies has been summarized diagrammatically in Figure 11.9. There have been four major achievements in the past year alone: (i) The use of pre-cast IPG technology, (ii) the first study on paralog gene identification using a genome-based strategy, (iii) the isolation of desired multiprotein complexes and identification of their protein components, and (iv) the generation of high-resolution 2D gel reference

1 Prepare the following solutions:

100 mM Ammonium bicarbonate (NH_4HCO_3), pH8.5 → 400 mg/50 mL MQ, add 120–140 μL 28% Ammonia solution
50% Acetonitrile/100 mM NH_4HCO_3
10 mM DTT(1.54 mg/mL)/100 mM NH_4HCO_3
50 mM Iodoacetamide(10.2 mg/mL)/100 mM NH_4HCO_3
1 $\mu\text{g}/\mu\text{L}$ trypsin in 20 mM NH_4HCO_3 : Store at -80°C → 0.01 $\mu\text{g}/\mu\text{L}$ trypsin in 20 mM NH_4HCO_3 (freshly diluted) ** 9 μL trypsin/720 μL 20 mM NH_4HCO_3 for 36 samples
50% Acetonitrile/0.5% Trifluoroacetic acid (or Acetic acid) → 750 μL Acetonitrile/750 μL MQ, add 7.5 μL of Trifluoroacetic acid

- 2 Dice each gel slice into small pieces (1 mm) and place into 1.5 mL TPX tubes.
- 3 Add 100 μL of **100 mM NH_4HCO_3** and keep for 5 min.
- 4 Using a gel loading pipet tip, extract the supernatant and discard.
- 5 Repeat steps 3 and 4 once or twice.
- 6 Add 200 μL of **50% Acetonitrile/100 mM NH_4HCO_3** and vortex for 10 min.
- 7 Remove supernatant, Speed Vac the gel pieces to complete dryness (~20 min).
- 8 Add 50 μL **10 mM DTT/100 mM NH_4HCO_3** to dried gels.
Vortex and spin briefly. Allow reaction to proceed at 56°C for 1 h.
- 9 Remove supernatant, add 50 μL **50 mM Iodoacetamide/100 mM NH_4HCO_3** to the gel pieces,
vortex, and spin briefly. Allow reaction to proceed in the dark for 45 min at room temperature.
- 10 Remove supernatant (discard). Wash gels with ~100 μL **100 mM NH_4HCO_3** , vortex for 10 min, spin.
- 11 Remove supernatant (discard). Dehydrate gels with 200 μL of Acetonitrile,
vortex for 5 min, and spin. Repeat one more time.
- 12 Speed Vac the gel pieces to complete dryness (~20 min). Proceed with trypsin digestion.
- 13 Add 20 μL **trypsin solution (0.01 $\mu\text{g}/\mu\text{L}$ trypsin in 20 mM NH_4HCO_3)**,
rehydrate the gel pieces on ice or at 4°C for 10 min.
- 14 Remove supernatant (discard).
- 15 Add 10 μL **20 mM NH_4HCO_3** as needed to cover the gel pieces,
spin briefly and incubate at 37°C for 4 h—overnight.
- 16 Transfer the digested solution (aqueous extract) into a clean 1.5 mL TPX tube.
- 17 To the gel pieces, add 30 μL (to cover) of **50% Acetonitrile/0.5% TFA**, vortex 20–30 min, spin,
sonicate for 5 min, and transfer digested solution to same TPX tube. Repeat one more time.
- 18 Speed Vac to reduce volume to 20–330 μL (~40 min), and store at -80°C .

➡ MS analysis

FIGURE 11.7. Standardized in-gel digestion protocol for silver nitrate stained protein spots. Differentially expressed spots are excised from the 2D gels using a gel picker (One Touch Spot Picker, P2D1.5 and 3.0, The Gel Company, San Francisco, CA) and stored at -30°C . Sequencing grade modified trypsin (17,000 U/mg; Promega, Madison, WI) is used for the digestion. The peptides in 20/30 μL solution can be used for mass spectral analysis on a LCQ Deca linear ion trap mass spectrometer (nESI–LC–MS/MS, Thermo Electron) through a nESI source.

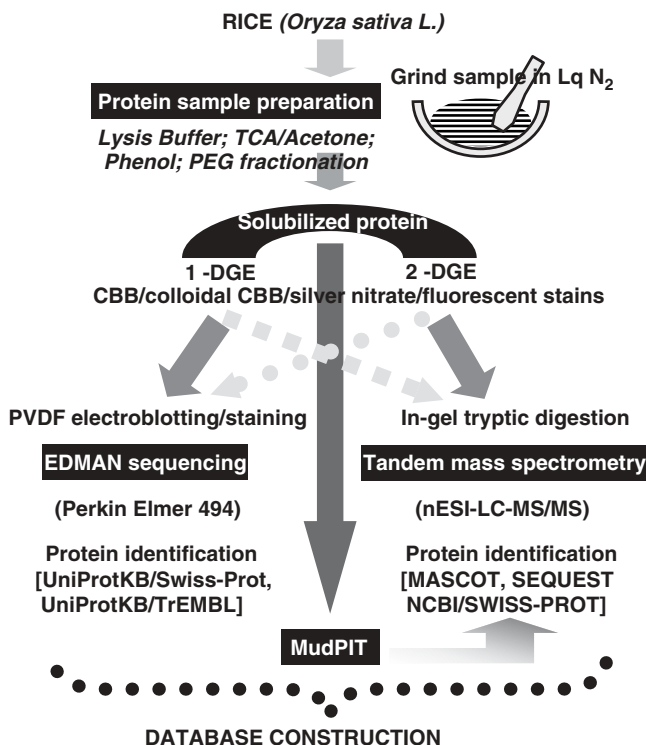


FIGURE 11.8. Simplified strategy for rice proteomics.

maps in many studies focusing on development (e.g., leaf and caryopsis organelles (etioplasts)), abiotic and biotic stresses (such as temperature and pathogens), disulfide proteins, lesion-mimic mutants (LMMs), hybrids, and transgenic plants. In particular, many investigations of protein complexes give a more direct and functional dimension to rice proteomics [4]. In the past year, many publications adhering to the use of pre-cast IPG strips for 2-DGE have been seen. While it is rewarding to see our predictions (on use of IPG) come true, much more progress needs to be achieved in terms of universal usage and acceptance.

Highlights of the Year 2007

Several well-conducted rice proteomics papers appeared 2007, and they are briefly reviewed here to get an idea where rice proteomics is heading.

The first paper we wish to discuss deals with pollen. Dai et al. [18] recently reported a new method for obtaining a large quantity of *in vitro* germinating rice pollen for proteomics studies. The rationale given by the authors is that after pollination, pollen germinates quickly, and methods for collecting large amounts of *in vitro*-germinated pollen grains for transcriptomics and proteomics studies were not

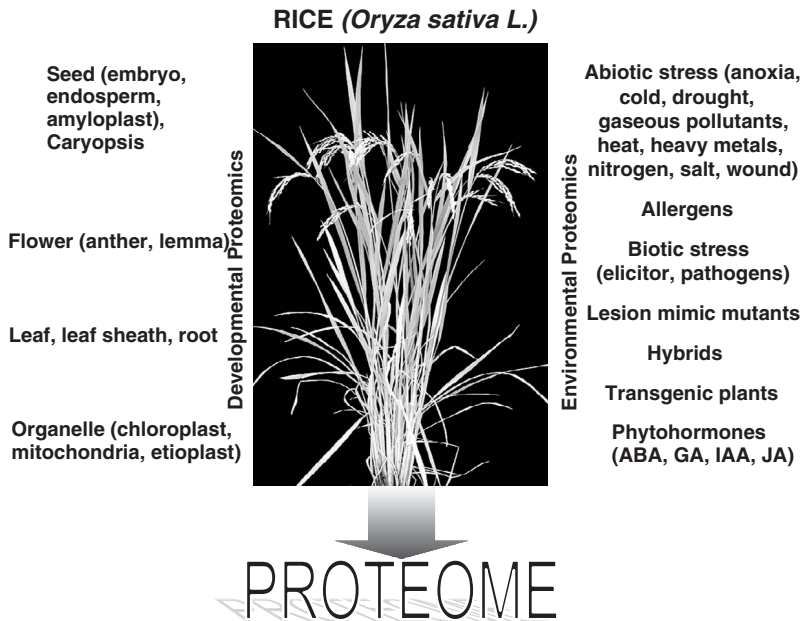


FIGURE 11.9. The rice proteomes.

previously available; therefore, molecular information about the germination developmental process was lacking [18]. Using 2-DGE, ~2300 protein spots were detected by the authors. Image analysis revealed that 186 proteins were differentially expressed in mature and germinated pollen. Among these, 66 proteins were developmentally regulated. MS analysis identified 120 diverse protein species involved in wall metabolism, protein synthesis and degradation, cytoskeleton dynamics, and carbohydrate/energy metabolism. This excellent study provides novel insights into pollen functional specialization in rice.

The second paper used two global approaches to develop an inventory of differentially expressed genes and proteins in rice (cv. Nipponbare) seedling grown on Murashige and Skoog medium, with and without jasmonic acid (JA) [19]. JA was found to significantly reduce the growth of shoots, roots, leaves, and leaf sheaths, depending on the concentration used ($1 < 2 < 5 < 10 < 25 < 50 \mu\text{M}$). Because $5 \mu\text{M}$ JA treatment showed almost 50% growth inhibition of seedlings, shoots and roots of these seedlings grown for 7 days were used for proteomics and transcriptomics (rice oligo DNA microarray, Agilent 22K) analyses. 2-DGE revealed 66 and 68 differentially expressed protein spots in the shoot and root, respectively, compared to controls. MS analysis of these protein spots identified 52 (shoot) and 56 (root) nonredundant proteins, belonging to 10 functional categories. Proteins involved in photosynthesis (44%), cellular respiratory (11%), and protein modification and chaperone (11%) were highly represented in shoots, whereas antioxidant system (18%), cellular respiratory

(17%), and defense-related proteins (15%) were highly represented in roots. Transcriptomics analysis of shoots and roots identified 107 and 325 induced genes and 34 and 213 suppressed genes in shoots and roots, respectively. Over 57% of the total gene count encoded for proteins involved in secondary metabolism, energy production, protein modification and chaperone, transporters, and cytochrome P450s. These identified proteins and genes were then discussed with respect to the JA-induced phenotype, providing a new insight into the role of JA in rice seedling growth and development [19].

The third paper deals with callus differentiation, a key developmental process for rice regeneration from cells [20], with the aim of gaining a better understanding of this complex developmental process. Yin et al. [20] used 2-DGE to reveal the temporal patterns of protein expression at early stages during rice callus differentiation. A total of 79 protein spots on 2D gels were analyzed by MS/MS, resulting in the identification of 60 proteins; most of which are known to play a role in plant development. Moreover, the authors identified two new proteins associated with the callus differentiation and further confirmed their findings by Western blot analysis. As in the preceding study, the results of proteomics experiments were verified at the mRNA level using microarray and real-time PCR. Although Cho et al. [19] did not find good correlation between protein expression and mRNA levels in their experiment, Yin et al. [20] showed a good correlation between differentially expressed protein levels with their corresponding mRNA levels. This study is an important contribution to understanding of the mechanisms that lead to callus differentiation of rice [20].

The fourth paper examines the rice spotted leaf 6 (*spl6*) mutant, which produces lesions caused by spontaneous cell death in the absence of pathogenic infection [21]. In this mutant, after the tillering stage, small red and brown lesions were initiated in groups on the leaf blade, subsequently forming parallel lines along the midrib of the leaf. Using light and transmission electron microscopy (TEM), thylakoid membranes of mesophyll chloroplasts were found to be progressively damaged in nonspotted section of mutant leaf. Conversely, in the spotted area, chloroplasts were found to be absent in mesophyll cells, suggesting that lesion formation in the *spl6* mutant might be due to an oxidative burst. 2-DGE analysis revealed 159 up- or down-regulated protein spots in a comparison between spotted (*spl6*) and normal leaves (wild-type). Protein identification results showed that protein disulfide isomerase (PDI), transketolase, thioredoxin peroxidase (TPX), ATP synthase, RuBisCO LSU, and RuBisCO activase were abundant in the wild-type but absent in the *spl6* mutant. These results prompt the authors to speculate on the cause of PCD and lesion development in the *spl6* mutant—that is, the failure to protect cells against oxidative burst by the absence of TPX and PDI in the leaves of the *spl6* mutant leaves [21].

The most recent rice study, at the time of writing, investigated a phenomenon that may be directly linked to global warming, namely, heat stress. Lee et al. [22] subjected rice seedlings to a temperature of 42°C, and it analyzed the leaves at 12 and 24 h after treatment. By observing increased ion leakage and lipid peroxidation, the authors suggested that oxidative stress is generated in rice leaves exposed to high temperature. 2-DGE in conjunction with MS was then used to identify heat-responsive proteins in rice leaves. Removal of RuBisCO using 15% PEG fractionation protocol [3] led to the identification of low-abundant proteins in leaves. Approximately 1000 protein spots

were reproducibly detected, where 73 protein spots were found to be differentially expressed at least at one time point (12 or 24 h). Thirty-four and 39 protein spots were found in the PEG supernatant and pellet fractions, respectively, from among the 73 protein spots. A total of 48 proteins were identified by MALDI-TOF-MS and categorized into classes related to heat shock proteins (HSPs), energy and metabolism, redox homeostasis, and regulatory proteins. A group of low molecular small HSPs (smHSPs) were found to be newly induced by heat stress. Among these, a mitochondrial smHSP protein expression was validated by immunoblot analysis. Lee and co-workers also examined the mRNA expressions of four differentially accumulated proteins that correspond to antioxidant enzymes but found that transcription levels were not completely concomitant with translation. Taken together, this study resulted in the identification of some novel proteins related to the heat stress response, providing new insights into the heat responsive proteome in rice.

11.4 CONCLUSIONS

Over the last three years, the number of publications on rice proteomics has increased substantially. It is pleasing to see so many new research groups devoting their energies into the study of the rice proteome. The progress seen in rice proteomics is mainly due to technology—from sample preparation to separation of proteins in the first and second dimensions, and to the analysis of the separated (or differentially expressed) proteins by MS. Furthermore, as visible in Figure 11.9, we have seen progress on the rice plant developmental aspects—the traditional approach of identifying proteins from rice seeds, flowers, leaf, shoot, and root. However, the effect of the environment on rice is of increasing interest and subject to many studies. Hybrids and transgenic plants are also being analyzed in order to provide an insight into their character. Indeed, rice proteomics has come to almost match the level of proteomics research in *Arabidopsis* and the gap is closing steadily. It should also be remembered that the initial research on *Arabidopsis* proteomes has given us valuable lessons and impetus for rice proteomics.

11.5 FIVE-YEAR VIEWPOINT

In the next five years, we expect (i) standardized protocols for protein extraction and separation based on the pre-cast IPG technology and, if possible, pre-cast SDS-PAGE, (ii) standardized protocols for subcellular proteomics, and (iii) detailed and high-quality 2D gel reference maps for almost all tissues and organelles. This will undoubtedly require advanced image analysis software and bioinformatics (see Chapter 8). MS analysis will also have to adhere to a high standard (see Chapters 3 and 9). Other than 2-DGE, which is and will remain the mainstay of proteomics, it will be useful to utilize 1-DGE coupled with LC-MS/MS to provide additional information of protein components, which may be missing from a standard 2-DGE-based approach. Broad utilization of MudPIT and 2D LC-based MS analysis will provide us with a vast

amount of proteomics data. Finally, quantitative proteomics demands attention. There is no doubt that the 2-DGE-based approach has given a useful quantitative analysis of proteins, and for this there are some good examples, such as that recently developed for seed filling in soybean [23] and rapeseed [24]. On the MS side, other than the expensive label-based quantitative approaches, label-free quantitative approaches will be very useful, economical, reproducible, and less prone to errors involved with the labeling reactions.

ACKNOWLEDGMENT

We are grateful to Ms. Junko Shibato for the preparation of the artwork.

REFERENCES

1. Smith BDT. 1998. *The Emergence of Agriculture*, Scientific American Library (Division of HPHLP), New York.
2. International Rice Genome Sequencing Project. 2005. *Nature* **436**:793–800.
3. Agrawal GK and Rakwal R. 2006. *Mass Spec Rev* **25**:1–53.
4. Agrawal GK, et al. 2006. *Proteomics* **6**:5517–5528.
5. *Arabidopsis* Genome Initiative. 2000. *Nature* **408**:796–815.
6. Bernal A, Ear U, and Kyrpides N. 2001. *Nucleic Acids Res* **29**:126–127.
7. Agrawal GK, et al. 2005. *J Chromatogr* **815**:109–123.
8. Agrawal GK, et al. 2005. *J Chromatogr* **815**:125–136.
9. Agrawal GK, et al. 2005. *J Chromatogr* **815**:137–145.
10. Jorin JV, Maldonado AM, and Castillejo MA. 2007. *Proteomics* **7**:2947–2962.
11. Rakwal R and Agrawal GK. 2003. *Electrophoresis* **24**:3378–3389.
12. O'Farrell PH. 1975. *J Biol Chem* **250**:4007–4021.
13. Cho K, et al. 2006. *J Plant Biol* **49**:413–420.
14. Zhan X and Desiderio DM. 2003. *Electrophoresis* **24**:1834–1846.
15. Edman P. 1949. *Arch Biochem Biophys* **22**:475–483.
16. Aebersold RH, et al. 1987. *Proc Natl Acad Sci USA* **84**:6970–6974.
17. Ouyang S, et al. 2007. *Nucleic Acids Res* **35** (database issue): D883–D887.
18. Dai S, et al. 2007. *Mol Cell Proteomics* **6**:207–230.
19. Cho K, et al. 2007. *J Proteome Res* **6**:3581–3603.
20. Yin L, et al. 2007. *Proteomics* **7**:755–768.
21. Kang SG, et al. 2007. *Proteomics* **7**:2447–2458.
22. Lee DG, et al. 2007. *Proteomics* **7**:3369–3383.
23. Hajduch M, et al. 2005. *Plant Physiol* **137**:1397–1419.
24. Hajduch M, et al. 2006. *Plant Physiol* **141**:32–46.

PROTEOMICS OF LEGUME PLANTS

Satish Nagaraj, Zhentian Lei, Bonnie Watson,
and Lloyd W. Sumner

12.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Proteomics has become an important complement to genetic data which is being used to extract novel functional information and to characterize plant systems. Several valuable reviews have now been published that focused upon the advancing status of plant proteome research [1–5], subcellular proteomics [6], and the importance of integrated genetic and proteomic approaches [7]. A review article by Barbier-Brygoo and Joyard [8] illustrated the key aspects concerning the rapid evolution and importance of proteomics in plant functional genomics. A comprehensive three-part review that discussed proteomics of dicot plants with regards to technologies [9], proteomes of developmental stages [10], and proteomes influenced by environment [11] provided extensive information on plant proteomics. Most proteomics studies to date in legumes have utilized 2-DGE and MS. The use of 2-DGE, though a relatively older approach, is still quite popular due to improvements in protein solubilization, enhanced resolution, and reproducibility [12]. The gel-based technologies, including 2-DGE, and MS have been dealt with in Chapters 2 and 3.

TABLE 12.1. Protein Extraction Protocol References for Different Legume Tissues/Organs

Plant System	Root	Leaf	Stem	Flowers	Seeds	Pods	Cultures
<i>Medicago truncatula</i>	21, 23, 29	23	23	23	24	23	23, 28
<i>Medicago sativa</i>	X ^a	49	49, 48	X	X	X	X
<i>Lotus japonicus</i>	35 ^b	X	X	X	X	X	X
<i>Glycine max</i>	44 ^b , 46 ^c	X	X	X	38, 39, 40	X	X

^aProtocol not available.

^bProteins isolated from peribacteroid membranes of root nodules.

^cProteins isolated from root hairs.

Legumes are a diverse plant family that includes approximately 18,000 species and the important agronomic crops soybean, alfalfa, pea, chickpea, and common bean. The phylogeny of legumes has been reviewed [13–16], and legumes are divided into three subfamilies: Mimosoideae, Caesalpinioideae, and Papilionoideae [16]. The Papilionoideae subfamily includes most of the economically important crop plants and differs from other legumes by several morphological traits that include orientation of the seed hilum and unidirectional initiation of sepals [15]. This subfamily consists of several clades, of which the phaseoloid and the galegoid clades contain all the major commercial crops with the exception of peanuts [13]. Cumulatively, the Papilionoideae subfamily contains five tribes, namely, Viceae, Trifolieae, Cicereae, Loteae, and Phaseoleae [13]. Most legume proteomics data generated to date have focused on the genera *Medicago*, *Lotus*, and *Glycine*. This chapter focuses on the proteomics of the genera *Medicago* (Trifolieae), *Lotus* (Loteae), and *Glycine* (Phaseoleae), with descriptions of specific organs/tissues proteomics of each group. Specific legume protein extraction protocols for a variety of organs/tissues are summarized with references and are provided in Table 12.1.

12.2 PROTEOMICS OF MODEL LEGUME *MEDICAGO TRUNCATULA*

Various traits including a small genome size, self-fertility and rapid generation time have promoted *M. truncatula* as a model plant for the legume family [17–19]. Genome sequencing of this model plant [20] and availability of over 225,000 ESTs (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) are immense assets for the molecular, genetic, and functional genomics study of legume biology. The following paragraphs provide a brief literature survey on *M. truncatula* proteomics. Figure 12.1 provides a snapshot of the tissues/organs of *M. truncatula* from which most of the proteome data has been generated. A 2D gel reference map has been established for *M. truncatula* root using an EST database and MALDI–TOF–MS PMF for protein identification [21]. The utility of this 2D gel reference map was then evaluated for cross-species protein identifications by the same group [22], whereby 2-DGE profiles of three legumes and a non-legume were analyzed. The authors reported a similarity in gel patterns between two bacterial strains and between the plants species. They also provided evidence to caution against using protein maps and PMF

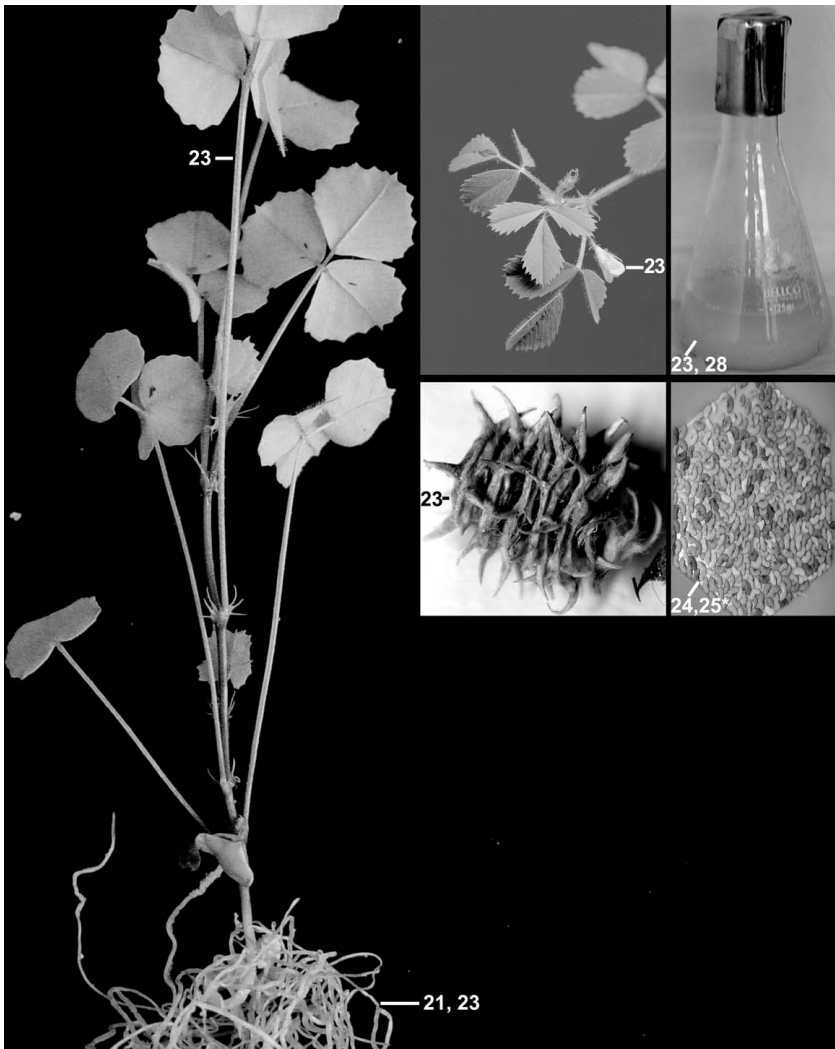


FIGURE 12.1. Proteomics snapshot of organs/tissues in *M. truncatula*. The numbers in the figures correspond to the numbers in the bibliography of this chapter. (Photographs courtesy of JiangQi Wen, Facility Manager, Molecular Biological Materials and IS Department of Noble Foundation).

for cross-species protein identification. 2D gel proteome reference maps for other *M. truncatula* organs/tissues have also been reported and included stems, roots, flowers, seed pods, and cell suspension cultures [23]. A protein identification success rate of approximately 55% was reported in this study, which identified 300 proteins by MALDI-TOF-MS and/or LC-MS/MS. This study also discussed and reported moderate correlation between the protein and mRNA levels. A detailed study of seed

development that investigated specific stages of seed filling identified about 84 proteins of the 120 proteins that yielded differential kinetics of appearance [24]. Furthermore, differentially accumulated proteins involved in carbon metabolism, embryonic photosynthesis, and methionine metabolism were discussed. The authors proposed that the proteins associated with carbon metabolism and embryonic photosynthesis may be involved in carbon dioxide (CO₂) refixation during seed filling or provide co-factors for protein and lipid synthesis and that the proteins of methionine metabolism may be involved in seed development. To further the understanding of seed development, the heat-stable protein fractions of desiccation-sensitive and desiccation-tolerant radicles were compared [25]. This analysis utilized the heat-stable proteome to enrich for LEA; and 2 of the 11 LEA proteins, identified as having a role in desiccation tolerance, were further characterized [25]. Two studies from the Rolfe group provided a comprehensive analyses of somatic embryogenesis in *M. truncatula* [26, 27]. This group has established proteome reference maps for embryogenic tissue culture cells and has also used them to analyze the protein profiles of globular somatic embryogenesis [26]. They identified 169 proteins belonging to various functional categories such as primary metabolism, hormone-regulated proteins, secondary metabolism, and plant defense. These reference maps are accessible by the scientific community at <http://semele.anu.edu.au/2d/2d.html>. Their more recent study compared tissue cultures from leaf explants of two different varieties (2HA and Jemalong) of *M. truncatula* which revealed significant changes in protein expression over time [27]. These changes might help in elucidating the molecular processes involved in somatic embryogenesis and also help in understanding the hyperembryogenesis of 2HA as compared to the wild-type Jemalong. Finally, an extensive 2D gel proteome reference map of *M. truncatula* cell suspension cultures has been reported [28]. This work led to the identification of 1367 proteins that will be useful in subsequent studies involving an integrated functional genomic study of response of cell suspension cultures to several biotic and abiotic stresses (Lei et al. and Nagaraj et al., unpublished data).

A major motivation for the study of legumes and hence the establishment of *M. truncatula* as a model system is the study of plant symbioses, which is not possible in *Arabidopsis*. In this regard, root protein profiles of *M. truncatula* that were inoculated with mycorrhizal fungi *G. mosseae* or nitrogen-fixing bacteria, *S. meliloti*, were studied [29]. This report cited the differential accumulation of several proteins and the induction of new proteins, both in mycorrhizal roots and nodulated roots, and their functional roles are discussed [29]. The authors reported the identification of proteins from plant and bacterial origin in the case of nodulated roots, but in mycorrhizal roots they identified only proteins of plant origin. Proteomics was also used to analyze the protein profiles of both root-colonizing and nodule-forming bacteria, *S. meliloti* [30], and its symbiotic host's symbiosome membrane protein profile [31]. Djordjevic's lab [30] compared protein profiles of bacteria harvested from nodules and cultured bacteria. They showed differences between the proteome of nodule bacteria and cultured bacteria. These differences indicated a shift in metabolism that occurs in the bacteria as a result of its interaction with the host plant. Sherrier's group [31] also examined the protein profile of the bacteria along with the symbiosome membrane protein profile.

Subcellular proteomics has been used to investigate the microsomal proteins of *M. truncatula* roots [32] and the response of the root membrane proteins to colonization of arbuscular mycorrhizae (AM) [33]. The former study contributed a fractionation technique for characterizing microsomal root proteins along with the identification of proteins with a success rate of over 80%. The latter provided information on proteins differentially regulated in response to fungal colonization. Also, they report for the first time the identification of a few fungal proteins in planta. Root adaptations to various stresses have been discussed in a recent study in which proteins induced in roots in response to a pathogen, *A. euteiches* were analyzed [34]. This study described the expression profiles of PR10 proteins along with up-regulation of a protein involved in secondary metabolism.

12.3 PROTEOMICS OF MODEL SYSTEM *LOTUS JAPONICUS*

The proteomics of peribacteroid membrane by Wienkoop and Saalbach appears to be the only current published work in *L. japonicus* [35]. Using an LC-MS/MS approach, the authors identified approximately 94 proteins that represent the largest number of proteins identified from peribacteroid membrane to date. Interestingly, these included a number of transporters and kinases that may have a role in interaction between the legume and rhizobia. Also, endomembrane proteins that might play a role in symbiosome biogenesis were also detected [35]. A recent review by Udvardi et al. [36] mentions ongoing research to investigate the proteomes of seed and root nodule development.

12.4 PROTEOMICS OF SOYBEAN, *GLYCINE MAX*

An initial 2-DGE proteomics study of soybeans investigated the heterogeneity of soybean seed proteins [37]. However, the majority of the polypeptides were tentatively identified as 7S and 11S globulins. A more recent work outlined the automation of high-throughput PMF for identifying proteins in soybean seeds [38]. The reported workflow was validated, and the method was utilized to analyze ~100 proteins that were initially resolved by 2-DGE. The authors suggested that this work could be a platform for future comparative proteomics in soybean seed-filling. The same group also utilized a high-throughput proteomics approach to study developmental stages of soybean seeds [39]. They reported the identifications and expression profiles for around 600 proteins. They also developed a web-based database (<http://oilseedproteomics.missouri.edu>) for storage and dissemination of this information to the scientific community. Another report involving methods development by Natarajan et al. [40] showed a comparison of protein extraction and solubilization methods for soybean seed proteins using 2-DGE and MS analysis. Their results suggest that the use of thiourea/urea and TCA is optimal for protein extraction and subsequent use in proteomic analysis.

The majority of soybean proteomics studies to date have concentrated on seeds to better understand their roles in nutrition and health. In this regard, Gianazza et al. [41]

analyzed soy proteins from seeds used in most clinical studies of hypercholesterolemia. Multiple samples used in Europe and the United States were analyzed by 2-DGE and MS [41]. The authors reported on the difference in soy proteins used in different clinical studies and discussed the implications of these results on the outcome of clinical studies. In another study, 422 proteins were identified in developing soybean seeds by 2-DGE and MS [42]. This research addressed the use of near-infrared reflectance spectroscopy in combination with chemical analysis for predicting protein and carbohydrate composition. Also, they reported a negative correlation between the seed storage protein G1 glycinin precursor and carbohydrates such as nonstarch polysaccharides and raffinose. As the first of its kind, recent research has applied comparative proteomics to analyze storage proteins in wild and cultivated soybean seeds [43]. The authors reported a higher degree of electrophoretic heterogeneity in the wild soybean seed storage proteins than its cultivated counterpart. They mentioned that comparative proteomics of this kind provided a better understanding of evolutionary relationship between genotypes and also helps in defining specific changes in protein composition that can occur in new soybean varieties.

Panther et al. [44] utilized proteomics to identify proteins associated with the peribacteroid membrane of soybean root nodules. In this work, about 31 proteins resolved by 2-DGE were excised and subjected to N-terminal sequencing for identification. The majority of the proteins identified were involved in protein processing; several proteins did not show similarity to any known proteins, suggesting the presence of novel proteins. Another study reported on the protein expression pattern of *B. japonicum*, which establishes a symbiosis with soybean [45]. In this work, a partial proteome map was generated and included the identifications of approximately 180 bacterial proteins [45]. This work paves the way for further research that will continue the characterization of important symbiotic proteins.

Comparative proteomics between roots and root hairs and between uninoculated and *B. japonicum*-inoculated roots have also been published [46]. This work reported the practicality of proteomic study of organisms without a complete-genome sequence. Subcellular proteomics has also been initiated whereby the protein profiles of soybean nodule and root mitochondria were compared [47]. This work presented the identities of several putative symbiosis-related proteins and discussed the physiological roles of differentially expressed proteins. The authors suggest that similar research to explore proteomes of other subcellular organelles would provide insight on metabolic compartmentalization in root nodules. Figure 12.2 summarizes some of the published literature on soybean proteomics.

12.5 PROTEOMICS OF ALFALFA, *MEDICAGO SATIVA*

Proteomics has also been utilized to study the widely cultivated crop, *M. sativa* [48, 49]. Specific extraction procedures for alfalfa stem cell wall proteins have been established and used to generate a reference map for cell walls consisting of more than 100 identified proteins [48]. Incamps et al. [49] have also published a proteome reference map of soluble proteins from alfalfa leaves and stems. This map was used to

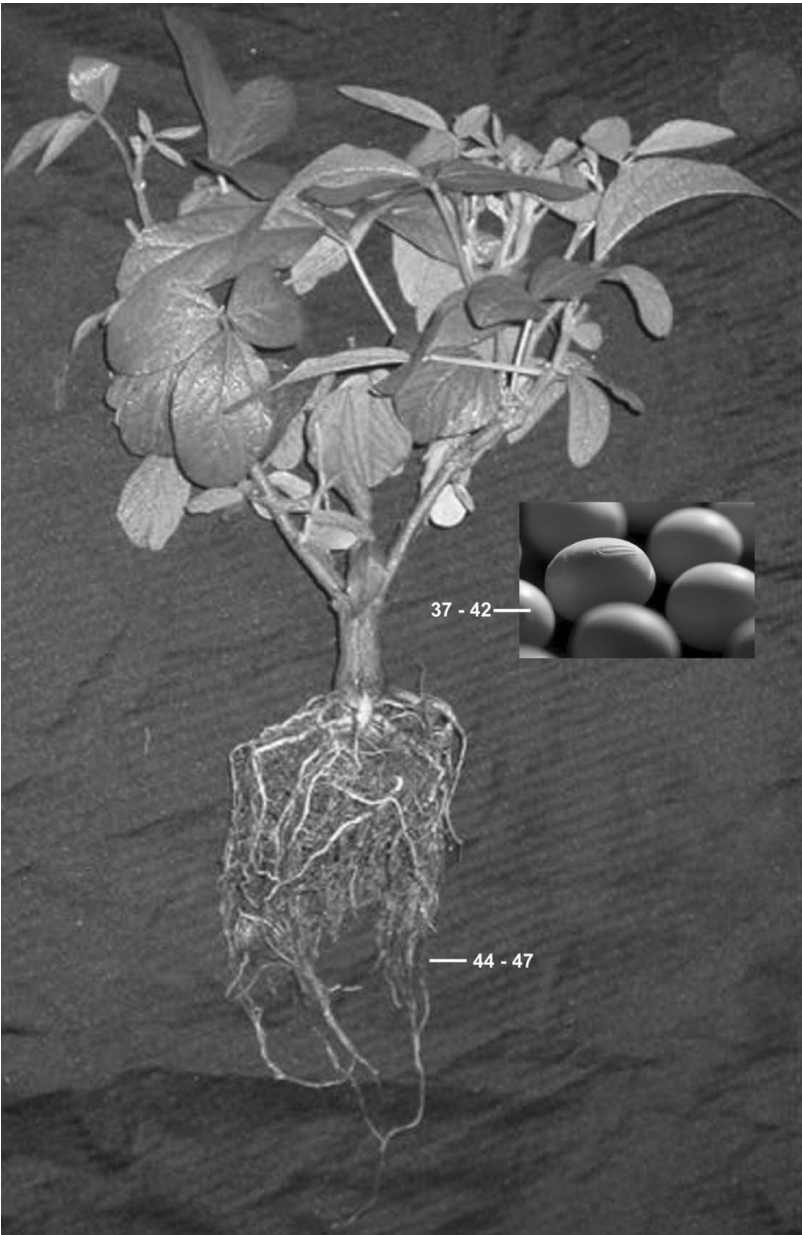


FIGURE 12.2. Proteomics snapshot of organs/tissues in *G. max*. The numbers in the figures correspond to the numbers in the bibliography of this chapter. (Photographs courtesy of Kristy Richerson, Greenhouse Associate, Noble Foundation).



FIGURE 12.3. Proteomics snapshot of organs/tissues in *M. sativa*. The numbers in the figures correspond to the numbers in the bibliography of this chapter. (Photographs courtesy of Kristy Richerson, Greenhouse Associate, Noble Foundation).

investigate the high protein content of products generated during wet fractionation of the green biomass of alfalfa. Figure 12.3 includes an overview of the proteome maps available in alfalfa.

12.6 CONCLUSIONS

The past five years have seen the inclusion of proteomics into the study of legumes. A major proportion of the efforts to date have focused on the development of critical infrastructure including methods development and the generation of proteome reference maps. However, it is believed that there are still significant opportunities for the use of proteomics to better understand legume biology particularly to answer certain species-specific questions in *L. japonicus* and other legume crops. Legume proteomics has been utilized to probe various aspects of legume biology including

plant organ differentiation, seed development, symbiotic interactions, and subcellular function. The majority of the work was accomplished using the legumes *M. truncatula*, *M. sativa*, and *G. max*. In *M. truncatula*, most proteins have been identified in cell cultures (~1400 proteins) [23, 28], followed by roots (~200 proteins) [21, 23] and seeds (~84 proteins) [24]. Approximately 260 proteins from leaf, stem, flowers, and seed pods, at an average of 53 proteins/organ or tissues, were identified in one study [23]. Only 94 proteins have been identified from the peribacteroid membranes isolated from *L. japonicus* [35]. A majority of proteins identified in *G. max* are from seeds (~200 proteins) [39], followed by another 80 proteins identified from roots [46, 47]. A total of 214 proteins have been identified from leaves and stems of *M. sativa* [48, 49]. These proteins from all organs/tissues of the above plants are the total number of proteins identified, including those obtained from comparative proteomics studies.

12.7 FIVE-YEAR VIEWPOINT

Proteomics will provide greater insight into post-translational regulatory mechanisms and protein interactions and complexes, neither of which are amenable to study at the molecular or genetic level. Furthermore, spatially resolved proteomics will enable a never-before-possible view of the anatomical segregation of plant biochemistry and enzymology. This information, in association with other functional genomics data including transcriptomics and metabolomics, offers great opportunities for empirically sound gene function annotation through the correlative analysis of co-localized transcripts, proteins, and metabolites. Finally, the current technologies utilized in modern proteomics are still somewhat limiting specifically in relation to comprehensive proteome coverage, dynamic range, hydrophobic protein analyses, and cost. Technical advances in any of these areas are expected to significantly elevate the utility of plant and legume proteomics. Particularly in legume proteomics, emphasis will continue on characterizing symbiosis. In-depth analysis that includes subcellular proteomics to understand symbiotic interactions and its effects on the organelles in plants will be important. Also, subcellular proteomics of organelles will increase our understanding of low abundant proteins. In addition, subcellular proteomics could be utilized for comparative analyses. Targeted studies directed toward applying proteomics to develop an interactome that represents interactions of proteins in a particular pathway of the plant could complement the information available in legumes. Lastly, phosphoproteomics and glycoproteomics, both of which are significant in understanding the PTMs, should be initiated to increase our understanding of legume biology.

REFERENCES

1. Canovas FM, et al. 2004. *Proteomics* **4**:285–298.
2. Thiellement H, et al. 1999. *Electrophoresis* **20**:2013–2026.
3. Zivy M and de Vienne D. 2000. *Plant Mol Biol* **44**:575–580.

4. van Wijk KJ 2001. *Plant Physiol* **126**:501–508.
5. Rossignol M. 2001. *Curr Opin Biotechnol* **12**:131–134.
6. van Wijk KJ. 2004. *Plant Physiol Biochem* **42**:963–977.
7. Thiellement H, Zivy M, and Plomion C. 2002. *J Chromatogr B Anal Technol Biomed Life Sci* **782**:137–149.
8. Barbier-Brygoo H and Joyard J. 2004. *Plant Physiol Biochem* **42**:913–917.
9. Agrawal GK, et al. 2005. *J Chromatogr B Anal Technol Biomed Life Sci* **815**:109–123.
10. Agrawal, GK, et al. 2005. *J Chromatogr B Anal Technol Biomed Life Sci* **815**:125–136.
11. Agrawal, GK, et al. 2005. *J Chromatogr B Anal Technol Biomed Life Sci* **815**:137–145.
12. Rabilloud, T. 2002. *Proteomics* **2**:3–10.
13. Choi H-K, et al. 2004. *Proc Natl Acad Sci USA* **101**:15289–15294.
14. Cronk Q, Ojeda I, and Pennington RT. 2006. *Curr Opin Plant Biol* **9**:99–103.
15. Doyle JJ and Luckow MA. 2003. *Plant Physiol* **131**:900–910.
16. Zhu H, et al. 2005. *Plant Physiol* **137**:1189–1196.
17. Trieu AT, et al. 2000. *Plant J* **22**:531–541.
18. Cook DR. 1999. *Curr Opin Plant Biol* **2**:301–304.
19. Bell CJ, et al. 2001. *Nucleic Acids Res* **29**:114–117.
20. Brink S, Chaffey N, and Stokes T. 2002. *Trends Plant Sci* **7**:100–102.
21. Mathesius U, et al. 2001. *Proteomics* **1**:1424–1440.
22. Mathesius U, et al. 2002. *Proteomics* **2**:1288–1303.
23. Watson BS, et al. 2003. *Plant Physiol* **131**:1104–1123.
24. Gallardo K, et al. 2003. *Plant Physiol* **133**:664–682.
25. Boudet J, et al. 2006. *Plant Physiol* **140**:1418–1436.
26. Imin N, et al. 2004. *Proteomics* **4**:1883–1896.
27. Imin N, et al. 2005. *Plant Physiol* **137**:1250–1260.
28. Lei Z, et al. 2005. *Mol Cell Proteomics* **4**:1812–1825.
29. Bestel-Corre G, et al. 2002. *Electrophoresis* **23**:122–137.
30. Djordjevic MA. 2004. *Proteomics* **4**:1859–1872.
31. Catalano CM, Lane WS, and Sherrier DJ. 2004. *Electrophoresis* **25**:519–531.
32. Valot B, Gianinazzi S, and Eliane DG. 2004. *Phytochemistry* **65**:1721–1732.
33. Valot B, et al. 2005. *Plant Mol Biol* **59**:565–580.
34. Colditz F, et al. 2004. *Plant Mol Biol* **55**:109–120.
35. Wienkoop S and Saalbach G. 2003. *Plant Physiol* **131**:1080–1090.
36. Udvardi MK, et al. 2005. *Trends Plant Sci* **10**:222–228.
37. Hu B and Esen A. 1982. *J Agric Food Chem* **30**:21–25.
38. Mooney BP and Thelen JJ. 2004. *Phytochemistry* **65**:1733–1744.
39. Hajduch M, et al. 2005. *Plant Physiol* **137**:1397–1419.
40. Natarajan S, et al. 2005. *Anal Biochem* **342**:214–220.
41. Gianazza E, et al. 2003. *J Nutr* **133**:9–14.
42. Hollung K, et al. 2005. *J Agric Food Chem* **53**:9112–9121.
43. Natarajan SS, et al. 2006. *J Agric Food Chem* **54**:3114–3120.
44. Panter S, et al. 2000. *Mol Plant Microbe Interact* **13**:325–333.

45. Sarma AD and Emerich DW. 2005. *Proteomics* **5**:4170–4184.
46. Wan J, et al. 2005. *Mol Plant Microbe Interact* **18**:458–467.
47. Hoa LT-P, et al. 2004. *Plant Cell Physiol* **45**:300–308.
48. Watson BS, et al. 2004. *Phytochemistry* **65**:1709–1720.
49. Incamps A, et al. 2005. *Biotechnol Bioeng* **91**:447–459.

PROTEOME OF SEED DEVELOPMENT AND GERMINATION

Julie Catusse, Loïc Rajjou, Claudette Job, and Dominique Job

13.1 INTRODUCTION

The seed is the major form of dissemination in plants. It results from the transformation of the fertilized ovule. Seed contains an embryo (the future plant), storage compounds necessary for the nutrition and growth of the embryo during germination (endosperm, cotyledons, perisperm), and teguments (a shield to protect the embryo against biotic and abiotic stresses). Seed development terminates, for most species growing in temperate climates, with a desiccation phase, after which the embryo enters in a quiescent state permitting its storage and survival for many years. The seed, therefore, occupies a central position in the plant life cycle. Mature seeds are resting organs, having low moisture content (5–15%) with metabolic activity almost at a standstill. In the absence of dormancy (a physiological process conditioning seed germination), for germination to occur, seeds need to be hydrated under conditions that encourage metabolism—for example, a suitable temperature and the presence of oxygen [1].

In angiosperms, seed formation results from a double fertilization process, where one of the sperm nuclei fuses with the egg nucleus to form the diploid zygote, while

a second sperm nucleus fuses with two polar nuclei giving rise to the triploid tissue endosperm. The embryo is often a plant in miniature possessing a root (radicle), a starting stem (hypocotyl), and leaves (cotyledons). By definition, the embryos of seeds of the dicots possess two cotyledons (e.g., *A. thaliana*, *M. truncatula*), whereas there is only one in monocots (e.g., rice) (Figure 13.1). Seeds can be categorized as endospermic or nonendospermic in relation to the presence or absence in the mature seed of a well-formed endosperm. In the latter case, other structures, usually cotyledons, are the principal storage organs, e.g., as in *Arabidopsis* (Figure 13.1). Two phytohormones, ABA and GAs, play key roles in controlling germination: The first molecule is a germination inhibitor involved in embryo development and maintenance of dormancy, while the second molecule promotes germination. With the completion of genome sequencing projects and the constitution of large EST collections for several reference and crop plants and the development of analytical methods for protein characterization, proteomics has become a major field of functional genomics. Broad proteomic analyses of seed development and germination have been initiated with the general aim of understanding the biochemical and molecular processes underlying seed quality and vigor.

13.2 PROTEOMICS OF DEVELOPING SEEDS

Several proteomics studies attempted to characterize the biosynthetic pathways responsible for accumulation of storage compounds in seeds. These reserves are of major importance for two reasons: (i) They support early seedling growth, and (ii) they are widely used for human and animal nutrition. Numerous biotechnological applications are expected toward improvement of nutritional value of seeds and for various nonedible uses (e.g., combustible fuel production or the use of seeds as a factory for the production/storage of recombinant proteins).

Dicot Seeds

Seeds of legume species are an important protein source, with 20% to as much as 40% protein content. However, the fact that the major proteins stored in these seeds are poor in sulfur-containing amino acids and the presence of nutritionally undesirable compounds, such as protease inhibitors, remain limiting factors. To address these questions, seed development in *M. truncatula* was investigated at specific stages of seed filling [2]. One hundred twenty proteins differing in kinetics of appearance were subjected to MALDI-TOF-MS. These analyses allowed us to identify 84 of them, some of which had previously been shown to accumulate during seed development in legumes (e.g., legumins, vicilins, convicilins, and lipoxygenases), confirming the validity of *M. truncatula* as a model for analysis of legume seed filling. The study also revealed proteins presumably involved in cell division during embryogenesis (tubulin and annexin). Their abundance decreased before the accumulation of the major storage protein families, which itself occurred in a specific temporal order: vicilins (14 days after pollination [DAP]), legumins (16 DAP), and convicilins (18 DAP). Furthermore, the study showed an accumulation of enzymes of carbon metabolism (e.g., sucrose synthase, starch synthase) and of proteins involved in embryonic

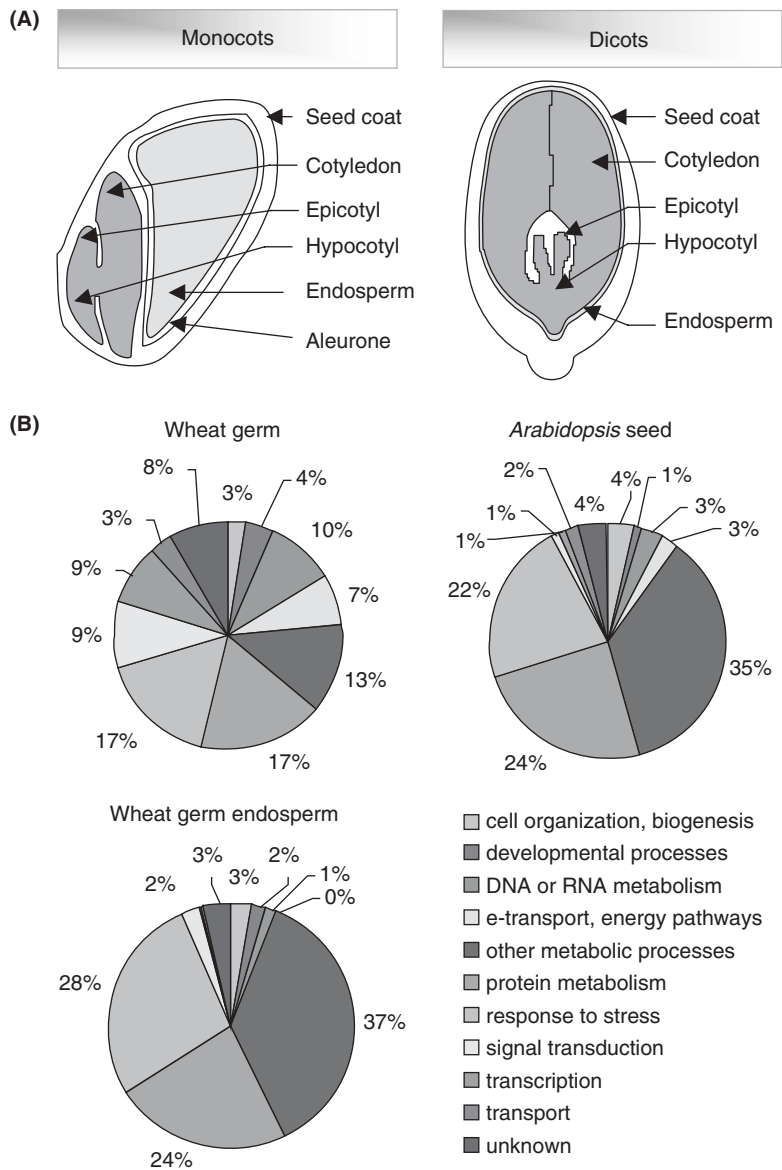


FIGURE 13.1. Structure of seeds and their proteomes. **(A)** Seed structure of monocots and dicots. **(B)** Albumin proteins from wheat germ (352 proteins analyzed; reference 13), mature *Arabidopsis* seeds (358 proteins analyzed; reference 20), and wheat endosperm (198 proteins analyzed; reference 7). See insert for color representation of this figure.

photosynthesis (e.g., chlorophyll *a/b* binding). Correlated with the reserve deposition phase was the accumulation of proteins associated with cell expansion (actin 7 and reversibly glycosylated polypeptide). Finally, this work revealed a differential accumulation of enzymes involved in Met metabolism [*S*-adenosylmethionine (AdoMet) synthetase and *S*-adenosylhomocysteine (AdoHcy) hydrolase] and proposed a role for these enzymes in the transition from a highly metabolically active to a quiescent state during seed development.

The expression profile and identity of hundreds of proteins during seed filling in soybean [3] have been analyzed at 2, 3, 4, 5, and 6 weeks after flowering (WAF) using 2-DGE and MS. This led to the establishment of high-resolution proteome reference maps (pH 4–7, 3–10) and quantitative expression profiles of 679 spots. These corresponded to 422 proteins representing 216 nonredundant proteins, which were classified into 14 major functional categories. Proteins involved in metabolism, protein destination and storage, metabolite transport, and disease/defense were the most abundant. To further detail global expression trends, composite expression profiles were established by summing protein abundance, expressed as relative volume, for each protein in each functional class for the five seed stages. Relative abundances of metabolic proteins decreased during the experimental period, suggesting metabolic activity curtails as seeds approach maturity (Figure 13.2). In contrast, the protein destination and storage class of proteins increased during late seed filling (Figure 13.2), and this was due to the preponderance of seed storage proteins.

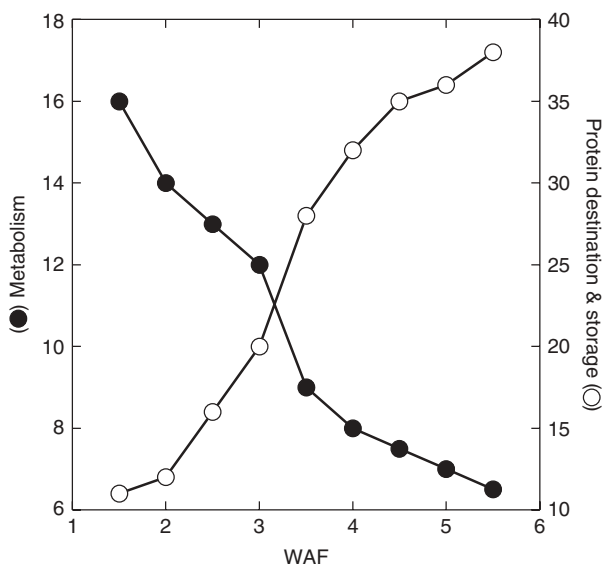


FIGURE 13.2. Regulation of metabolism during soybean seed filling (from reference 3). Combined expression profiles of all 82 proteins involved in metabolism (●) and of 80 proteins involved in protein destination and storage (○) were calculated as the sum of all relative volumes for each protein in several functional categories. WAF, weeks after flowering.

B. napus (also known as rape and oilseed rape) is the third largest oilseed crop in the world, providing approximately 13% of the world's supply of vegetable oil. *B. napus* seeds also produce proteins as main storage compounds. To characterize the biosynthetic pathways responsible for accumulation of these storage components, *B. napus* seed proteins were analyzed at 2, 3, 4, 5, and 6 WAF using 2-DGE and MALDI-TOF-MS/LC-MS/MS [4]. Developmental expression profiles for 794 protein spot groups were established, and hierarchical cluster analysis revealed 12 different expression trends. The identity of 517 spot groups was determined, representing 289 nonredundant proteins. These proteins were classified into 14 functional categories. Energy- and metabolism-related proteins were highly represented in developing seed, accounting for 24.3% and 16.8% of the total proteins, respectively. Analysis of subclasses within the metabolism group revealed coordinated expression during seed filling. These data provided an in-depth proteomic view on carbon assimilation in *B. napus* seed. In particular, they suggested that sugar mobilization from glucose to coenzyme A and its acyl derivative involves cooperation between the cytosol and plastids and that temporal control of enzymes and pathways extends beyond transcription. Also, as for soybean [3], relative abundances of metabolic proteins decreased during late stages of development as seeds entered into a quiescent state [4].

Monocot Seeds

Proteomics approaches were also used to investigate embryo development in monocots. Expressions of more than 400 polypeptide spots during rice caryopsis development were monitored [5]. Among them, more than 70 were characterized by MS, allowing the identification of 54 proteins with known functions. Of these, 21 were involved with carbohydrate metabolism, 14 with protein synthesis and sorting, including the major storage proteins glutelins and prolamins, and nine with stress responses, most of them being HSPs. The data indicate that changes in the expression of these proteins can be used to mark the physiological development stage of rice caryopses. Waxy (Wx) proteins and glutelins were the most significant spots, which increased significantly during development. Expression of large isoforms of Wx proteins was correlated with the amylose content of rice caryopses. High temperature (35/30°C) decreased the expression of Wx proteins, allergen-like proteins, and elongation factor (EF) 1b, but increased the expression of smHSPs, GAPDH, and prolamin. smHSP accumulation was positively correlated with the appearance of chalky kernels. During development, four Wx proteins isoforms were phosphorylated, while several glutelins were phosphorylated and glycosylated.

2-DGE was used for a time-resolved study of the changes in proteins that occur during seed development in barley [6]. Thirty-six selected spots were identified by MALDI-TOF-MS or by nESI-LC-MS/MS. Some proteins were present throughout development (cytosolic malate dehydrogenase, MDH), whereas others were associated with the early grain filling (APX), desiccation (cold-regulated protein Cor14b), or late (embryo-specific protein) stages. Most noticeably, the development process was characterized by an accumulation of serpin, low- M_r α -amylase inhibitors (presumed to defend the starch reserves of the seed against invading insect pathogens), Ser protease

inhibitors, and enzymes involved in protection against oxidative stress (glyoxalase I, APX, 1cys-peroxyredoxin). The expression of this last class of proteins throughout the development process most presumably reflected the importance of protection against ROS produced during seed development. The study also presented examples of proteins not previously experimentally observed (e.g., Cor14b), differential extractability of thiol-bound proteins, and possible allele-specific spot variation (e.g., at the level of β -amylase).

A proteomic approach was utilized to identify over 250 proteins of developing wheat starchy endosperm that participate in 13 biochemical processes [7]. Endosperm protein populations were compared at early (10 days post-anthesis, DPA) and late (36 DPA) stages of grain development. Analysis of protein number and spot volume revealed that carbohydrate metabolism, transcription/translation, and protein synthesis/assembly were the principal endosperm functions at 10 DPA. In marked contrast, stress/defense and storage were predominant processes at 36 DPA. These specific features are consistent with the ultimate role of the mature grain in reproduction. A major progress in the characterization of metabolic activity in developing wheat endosperm was achieved by proteomics of amyloplasts isolated from the starchy endosperm of developing wheat seeds (10 DPA) [8]. This study led to the identification of 289 proteins that function in a range of processes. The results highlighted the role of the amyloplast as a starch-storing organelle that fulfills a spectrum of biosynthetic needs of the parent tissue. When compared with a recent proteomic analysis of whole endosperm [7], this study demonstrated the advantage of using isolated organelles in proteomic studies.

A proteome reference map for developing maize endosperm has been established by means of 2-DGE and protein identification with LC-MS/MS analysis [9]. Among the 632 protein spots processed, 496 were identified: 42% against maize sequences, 23% against rice sequences, and 21% against *Arabidopsis* sequences. Identified proteins were not only cytoplasmic but also nuclear, mitochondrial, or amyloplastic. Metabolic processes, protein destination, protein synthesis, cell rescue, defense, cell death, and aging are the most abundant functional categories, comprising almost half of the 632 proteins analyzed in that study. This proteome map constitutes a powerful tool for physiological studies and is the first step for investigating the maize endosperm development.

13.3 PROTEOMICS OF MATURE SEEDS

The systematic study of plant proteins dates from the 19th century and in particular from the work of Osborne [10], who introduced the widely used classification into solubility groups based on their sequential extraction in water (albumins), dilute salt solutions (globulins), alcohol-water mixtures (prolamins), and dilute acid or alkali solutions (glutelins).

Globulins

The 11S–12S globulins are abundant seed storage proteins, being widely distributed in higher plants. They are synthesized during seed maturation in a precursor form of

about 60 kDa. At later stages, the precursor form is cleaved, yielding the mature globulins generally found in mature seeds. These are composed of six subunit pairs that interact noncovalently, each of which consists of an acidic A-subunit of $M_r \approx 40,000$ and a basic B-subunit of $M_r \approx 20,000$ covalently joined by a single S–S group. They are subsequently broken down during germination and used by the germinating seedling as an initial food source. A proteomics approach revealed that mature *Arabidopsis* seeds contain three forms of 12S globulins (cruciferins): (a) residual precursor forms, (b) A- and B-subunits, and (c) proteolyzed forms of A- and B-subunits [11]. These features suggest that (i) the maturation process giving rise to the formation of the A- and B-chains was not fully completed when developing seeds entered into quiescence and (ii) an early mobilization of the cruciferins occurred during the maturation phase. Thus, this study revealed that the anabolic processes that occur before germination and the catabolic processes that normally occur during germination are not fully separated developmentally in *Arabidopsis*. The role of specific proteases, the seed-type members of the vacuolar processing enzyme (VPE), in seed protein processing during seed filling and maturation has been investigated by proteomics, using knockout mutant alleles of all four members (αVPE , βVPE , γVPE , and δVPE) of the VPE gene family in *Arabidopsis* [12]. The complete removal of VPE function in the quadruple mutant resulted in a total shift of storage protein accumulation from wild-type processed polypeptides to a finite number of prominent alternatively processed polypeptides cleaved at sites other than the conserved asparagine residues targeted by VPE. Although these polypeptides largely accumulated as intrasubunit S–S-linked polypeptides, they showed markedly altered solubility and protein assembly characteristics. Instead of forming 11S hexamers, they were deposited primarily as 9S complexes. However, despite the impact on seed protein processing, plants devoid of all known functional VPE genes appeared unchanged with regard to protein content in mature seeds, relative mobilization rates of protein reserves during germination, and vegetative growth. These findings indicated that VPE-mediated Asn-specific proteolytic processing, along with the physiochemical property changes attributed to this specific processing step, is not required for the successful deposition and mobilization of seed storage protein in the protein storage vacuoles of *Arabidopsis* seeds.

Albumins

In our study of the albumin fraction of *Arabidopsis* seed we have now characterized 359 proteins, the major part of which can be assigned to the following functions: metabolic processes (35%), protein metabolism (24%), and response to stress (22%) (Figure 13.1). This representation is radically different from that of wheat (a monocot species) germ (Figure 13.1) [13]. Here, identified proteins were spread into a number of functional classes associated with processes that will be involved during germination in activating growth and development of the embryo such as transcription, translation, energy and general metabolism, transport, cell division, and signaling processes (Figure 13.1). It is striking that the proteome of wheat endosperm [7] is very different from that of the germ (Figure 13.1). A comparison of these data suggests that the protein functions present in mature dicot embryos are split between the germ and the endosperm in monocot seeds.

13.4 PROTEOMICS OF GERMINATING SEEDS

Seed germination is a complex, multistage process that can be divided into three phases—imbibition, increased metabolic activity, and initiation of growth—which loosely parallel the triphasic water uptake of dry mature seeds. Morphologically, initiation of growth corresponds to radicle emergence; subsequent growth is generally defined as seedling growth. By definition, germination *sensu stricto* incorporates those events that start with the uptake of water by the nondormant quiescent dry seed and terminate with the protrusion of the radicle and the elongation of the embryonic axis. Upon imbibition, the quiescent dry seed rapidly resumes metabolic activity. Recent studies addressed the question of the exact requirements for germination, particularly in terms of *de novo* RNA and protein syntheses.

A proteomic study of *Arabidopsis* seed germination showed that proteins associated with germination *sensu stricto* correlated with initial events in the mobilization of protein and lipid reserves and with the resumption of cell cycle activity, such as WD-40 repeat protein, tubulin, and cytosolic GAPDH [11]. During radicle emergence, proteins mostly involved in defense mechanisms to protect the future seedlings against herbivores, pathogens, and other stresses were identified, such as myrosinase, jasmonate-induced myrosinase-binding proteins, LEA, and HSP70 proteins. Among a total of 19 imbibition-associated proteins, seven proteins were identified, including actin 7 (ACT 7) and WD-40 repeat proteins.

The role of GAs in germination of *Arabidopsis* seeds was investigated using a GA-deficient *gal* mutant and wild-type seeds treated with paclobutrazol, a specific GA biosynthesis inhibitor [14]. As expected, with both systems, radicle protrusion was strictly dependent on exogenous GAs. The proteomic analysis revealed that GAs do not participate in many processes involved in germination *sensu stricto*—that is, the initial mobilization of seed protein and lipid reserves. Changes in 46 proteins were detected at this stage in the wild-type and mutant seeds. However, only one protein (α -2,4-tubulin) was suggested to depend on the action of GA, because it was not detected in the *gal* mutant seeds and accumulated in the mutant seeds incubated in the presence of GAs. In contrast, it was suggested that GAs might be involved, directly or indirectly, in controlling the abundance of several proteins (two isoforms of AdoMet synthetase and β -glucosidase) associated with radicle protrusion and postgermination processes. This proteomic study established for the first time the developmental stage at which GAs exert their action during germination, namely the radicle protrusion step, and unraveled several protein targets that can account for its action at this step.

The mechanisms controlling seed dormancy in *Arabidopsis* have been characterized by proteomics using the dormant accession Cvi originating from the Cape Verde Islands [15]. Comparative studies carried out with freshly harvested dormant (D) and after-ripened nondormant (ND) seeds revealed a specific differential accumulation of 32 proteins. The data suggested that proteins associated with metabolic functions potentially involved in germination can accumulate during after-ripening in the dry state, leading to dormancy release. Exogenous application of ABA to ND seeds strongly impeded their germination, which physiologically mimicked the behavior of D imbibed seeds. This application resulted in an alteration of the accumulation pattern

of 71 proteins. There was a strong down-accumulation of a major part (90%) of these proteins, which were involved mainly in energetic and protein metabolisms. This feature suggested that exogenous ABA triggers proteolytic mechanisms in imbibed seeds. An analysis of *de novo* protein synthesis by 2-DGE in the presence of [³⁵S]-Met disclosed that exogenous ABA does not impede protein biosynthesis during imbibition. Furthermore, imbibed D seeds proved competent for *de novo* protein synthesis, demonstrating that impediment of protein translation were not the cause of the observed block of seed germination. However, the 2D protein profiles were markedly different from those obtained with the ND seeds imbibed in ABA. Altogether, the data showed that the mechanisms blocking germination of the ND seeds by ABA application are different from those preventing germination of the D seeds imbibed in basal medium.

Proteomics was used to identify major proteins in extracts of germinating barley seeds [16]. This resulted in 198 identifications of 103 proteins in 177 spots from the mature barley seeds. These included housekeeping enzymes (glycolysis, starch metabolism, and citric acid cycle), chaperones, defense proteins (including enzyme inhibitors presumably involved in defense of the germinating seed against pathogens), and proteins related to desiccation and oxidative stress. Numerous spots in the 2D gel pattern changed during germination (micromalting), and an intensively stained area that contained large amounts of the serpin (serine protease inhibitor) protein Z appeared centrally on the 2D gel. Spots containing α -amylase also appeared. Identification of 22 spots after three days of germination represented 13 different database entries and 11 functions including hydrolytic enzymes, chaperones, housekeeping enzymes, and inhibitors. Despite the fact that cereal seed proteins have been studied for decades, new proteins have been identified on 2D gels. Several of the identified proteins have homologues in other organisms, but their function in barley seeds is not clear. Yet, their identification will facilitate the analysis of the changes in the proteome that occur during seed development and germination of cereal grains.

Two PTMs of seed proteins have been documented during germination. The first deals with the role of the NADP/thioredoxin system, composed of NADPH, thioredoxin *h*, and NADP-thioredoxin reductase in cereal seed germination. This system was shown to function in the reduction of the major storage proteins of the wheat endosperm, gliadins, and glutenins, converting S–S bonds to the SH state during germination [17]. These studies revealed new functions for thioredoxin in seeds, and they conclusively documented the general occurrence of a sequence of redox changes taking place in cereal endosperm: Proteins are synthesized in the reduced state early in seed development and oxidized during maturation and drying. Upon germination, thioredoxin reduces the oxidized proteins, thereby leading to increased solubility, proteolysis, and, ultimately, nitrogen and carbon mobilization. Another protein modification concerns the oxidation (carbonylation) of *Arabidopsis* seed proteins. A proteomics investigation showed that protein carbonylation can be detected in dry mature seeds and during the first stages of germination, which was a direct demonstration of the accumulation of ROS in seed development and germination [18]. In the dry mature seeds, the 12S cruciferins were the major targets. During imbibition, various carbonylated proteins accumulated. This oxidation damage was not evenly distributed among seed proteins but targeted specific proteins (e.g., glycolytic enzymes). Although

accumulation of carbonylated proteins is usually considered in the context of aging in a variety of model systems, this was clearly not the case for the *Arabidopsis* seeds because they germinated at a high rate and yielded vigorous plantlets. Thus, the results supported the hypothesis that the observed specific changes in protein carbonylation patterns are probably required for counteracting and/or utilizing the production of ROS caused by recovery of metabolic activity in the germinating seeds. In particular, the results suggested that blocking glycolysis could be beneficial during conditions of oxidative stress because it would result in an increased flux of glucose equivalents through the pentose phosphate pathway, thus leading to the generation of NADPH. This could in turn provide the reducing power for antioxidant enzymes, including the thioredoxin and GSH/glutaredoxin systems. Hence, protein carbonylation provides a means to adapt embryo metabolism to the oxidative conditions encountered during germination. Since carbonylation of proteins increases their susceptibility to proteolytic cleavage, the data also suggest that carbonylation of 12S cruciferins occurring during seed development facilitates their mobilization during germination, which might be advantageous for seedling establishment.

The defense response of plants against fungal infection has been intensively studied in vegetative organs, but quite surprisingly not in seeds. This question was addressed for the first time in germinating maize embryos by a proteomics approach [19]. The data highlighted the unexpected occurrence of a defense response during germination, which most presumably is required for protection of the germinating seed and the establishment of a vigorous plantlet. The influence of SA on elicitation of defense mechanisms in *Arabidopsis* seeds and seedlings was assessed by proteomics [20]. These analyses disclosed a specific role of SA in enhancing the reinduction of the late maturation program during early stages of germination, thereby allowing the germinating seeds to reinforce their capacity to mount adaptive responses in environmental water stress. Other processes affected by SA concerned the quality of protein translation, the priming of seed metabolism, the synthesis of antioxidant enzymes, and the mobilization of seed storage proteins. All these effects are likely to improve seed vigor. Another aspect revealed by this study concerned the oxidative stress entailed by SA in germinating seeds, as inferred from a characterization of the carbonylated proteome. Finally, the proteomics data revealed a close interplay between ABA signaling and SA elicitation of seed vigor.

A characteristic feature of seeds is their remarkable tolerance to dehydration stress. Soon after germination, this tolerance is lost. To understand the molecular basis of such desiccation tolerance, a proteomic analysis was performed on the heat-stable protein fraction of imbibed radicles of *M. truncatula* seeds [21]. Radicles were compared before and after emergence (2.8 mm long) in association with the loss of desiccation tolerance, as well as after reinduction of this tolerance by an osmotic treatment. The abundance of 15 polypeptides was linked with desiccation tolerance, out of which 11 were identified as LEA proteins. Interestingly, the data suggested that the LEA proteins expressed in seeds can be divided in two groups: (1) those that are induced only in tissues that are desiccation tolerant and (2) those that are also induced in osmotically shocked radicles that remain desiccation sensitive but do increase their tolerance to drying.

Mature dry seeds contain mRNAs stored during maturation. To characterize their role, the effect of α -amanitin, a transcriptional inhibitor targeting DNA-dependent RNA polymerase II, was examined on the germination of *Arabidopsis* seeds [22]. Germination (radicle protrusion) was shown to occur in the absence of transcription, while seedling growth was blocked. In contrast, germination was abolished in the presence of the translational inhibitor cycloheximide. Taken together, the results highlighted the role of stored proteins and mRNAs for germination in *Arabidopsis* and showed that in this species the potential for germination is largely programmed during the seed maturation process. To characterize the *de novo* synthesized proteome during germination, proteomics studies were carried out with this system, in the presence of [^{35}S]-Met as a labeled precursor (Figure 13.3). Proteins for which *de novo* synthesis was repressed by α -amanitin are involved in reactivation of metabolic activity during germination—for example, mitochondrial enzymes as the mitochondrial processing peptidase and succinate dehydrogenase, methionine biosynthesis, triacylglycerol metabolism, and hexose assimilation, a finding that emphasizes the importance of metabolic control in seed germination. Interestingly, a comparison of silver-nitrate-stained gels and radioactive gels revealed a number of spots whose levels apparently remained constant during germination, although the proteins were labeled with [^{35}S]-Met, thereby implying the occurrence of protein turnover. This finding revealed the existence of regulatory mechanisms to maintain constant the accumulation levels of some proteins during germination. More generally, these data illustrate the power of combining classical proteomics with dynamic proteomics in the interpretation of protein accumulation patterns. It must be stressed that the amount of a protein spot in a 2D gel reflects the accumulation level of that protein and not its rate of synthesis. The use of labeled precursors of protein synthesis is necessary to clarify this question. This distinction between accumulated and *de novo* synthesized proteins is also important in the perspective of comparing transcriptomics and proteomics expression data.

13.5 PROTEOMICS OF SOMATIC EMBRYOGENESIS

Somatic embryogenesis is a process analogous to zygotic embryogenesis, in which a single cell or a small group of vegetative (i.e., somatic) cells are the precursors of the embryos. This phenomenon can be divided into four major steps: (i) initiation of proembryogenic masses, (ii) proliferation of embryogenic cultures, (iii) maturation of somatic embryos, and (iv) regeneration of whole plants. Somatic embryogenesis is widely investigated in several plants because, on one hand, it provides useful systems for plant propagation (e.g., conifer biotechnology for reforestation programs) and, on the other hand, it allows fundamental studies on embryo development.

A proteomics study used leaf explants from the mutant line 2HA of *M. truncatula*, which presents a 500-fold greater capacity to regenerate plants in culture by somatic embryogenesis than the wild-type Jemalong cultivar chosen as reference for genomic studies [23]. Both 2HA and Jemalong leaf explants were grown on media containing the auxin 1-naphthaleneacetic acid and the cytokinin 6-benzylaminopurine. Proteins were extracted at different time points (2, 5, and 8 weeks) and were analyzed

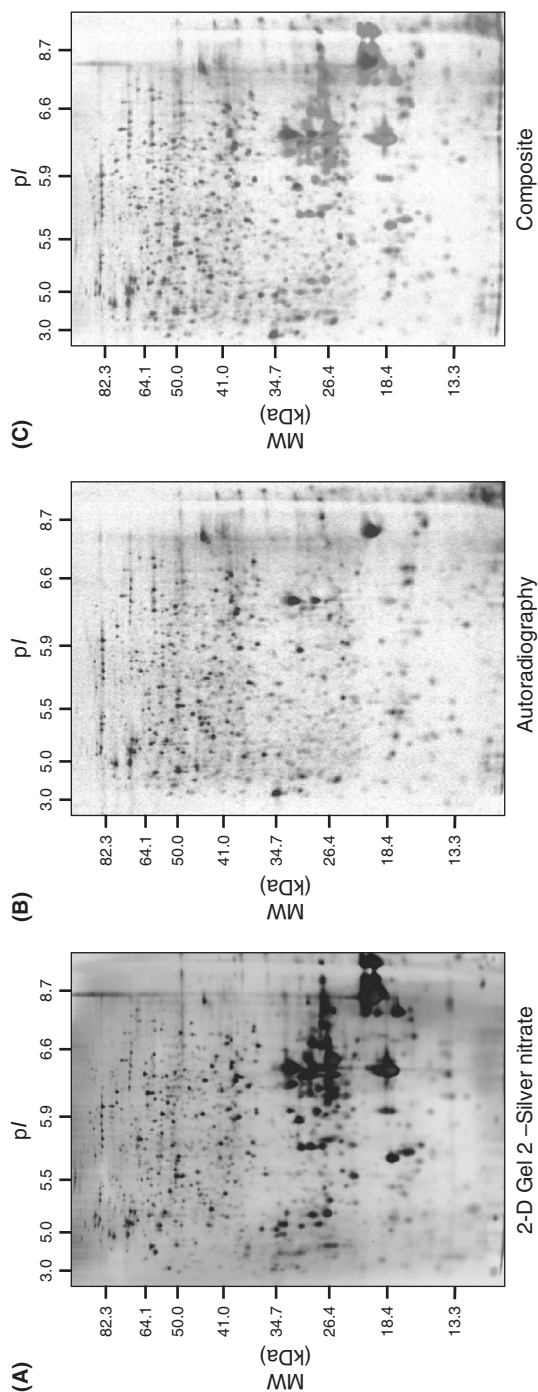


FIGURE 13.3. De novo protein synthesis during germination of *Arabidopsis* seeds (from Job et al., unpublished results). Seeds were incubated for 24 h in the presence of [35 S]-Met. Proteins were extracted and then submitted to 2-DGE, and the radiolabeled proteins were revealed by phosphorimager analysis. **(A)** Silver-stained gel. **(B)** Radiolabeled proteins from same gel as in A. **(C)** Superimposition of 2D patterns shown in A and B using false colors. Spots in red, proteins detected only by autoradiography; spots in green, proteins detected only by silver staining; spots in blue, proteins detected both by silver staining and autoradiography. See insert for color representations of this figure.

by 2-DGE. Fifty-four protein spots were significantly changed in accumulation during the 8 weeks of culture, 16 of these spots were identified by MALDI-TOF-MS or LC-MS/MS analyses. More than 60% of them had very different patterns of accumulation between 2HA and Jemalong during the 8 weeks of culture. Among the identified proteins, RuBisCO SSU proteins were gradually decreased in both Jemalong and 2HA during explant cultures. As such, this trend can be used as a marker for dedifferentiation and proliferation of the mesophyll tissues. Two of the most abundant proteins were an ABA-responsive protein with homology to PR10-1 and PR10-1 itself in both Jemalong and 2HA. They were not detected in the young leaves from which the explant cultures originated. Interestingly, they changed little throughout the 8 weeks of culture, suggesting a general role for ABA-responsive proteins and PR10 proteins in cell maintenance or cell defense. Chaperone proteins (dnaK-type HSP70 and luminal binding) showed a decrease in the 8-week-old cultures in both Jemalong and 2HA, although their accumulation levels were different. This may imply that a higher level of expression of the chaperones is required for the maintenance of cells during early culture. This study also identified proteins involved in seed formation (a seed maturation protein and a vicilin) as being expressed only in the highly embryogenic 2HA of 8-week-old cultures, testifying that somatic embryogenesis closely resembles zygotic embryogenesis. One of the most interesting proteins identified was thioredoxin *h*, suggesting the importance of redox regulation during commitment from the vegetative stage to a pathway of cellular differentiation and proliferation.

A proteomic approach was employed to quantitatively assess the expression levels of proteins across four stages of somatic embryo maturation in white spruce [24]. Forty-eight differentially expressed proteins have been identified by LC-MS/MS. These proteins are involved in a variety of cellular processes, many of which have not previously been associated with embryo development. It must be stressed that the combined use of the spruce ESTs in conjunction with GenBank accessions for other plants improved the rate of protein identification from 38% to 62%. This underscored the utility of EST resources in a proteomic study of any species for which a genome sequence is unavailable. Interestingly, as in zygotic embryogenesis (see above), AdoMet synthetase was detected at high level in the immature embryos and decreased to background level in mature embryos, suggesting common features of metabolic regulation in somatic and zygotic embryogenesis. Furthermore, as in zygotic embryogenesis, developing somatic embryos proved capable of accumulating storage proteins as vicilins.

13.6 METABOLIC CONTROL OF SEED DEVELOPMENT AND GERMINATION

Among the essential amino acids synthesized by plants, methionine is a fundamental metabolite because it functions both as a building block for protein and as the precursor of AdoMet, the universal methyl-group donor and the precursor of polyamines and the plant-ripening hormone ethylene. During *Arabidopsis* seed germination, two enzymes in this pathway showed differential accumulation [25]. The first corresponded

to methionine synthase. This protein was present at low level in dry mature seeds, and its level was increased strongly at 1-day imbibition, prior to radicle emergence. Its level was not increased further at 2-day imbibition, coincident with radicle emergence. The second enzyme corresponded to AdoMet synthetase, which was detected in the form of two isozymes with different pI and M_r . Both proteins were absent in dry mature seeds and in 1-day imbibed seeds but specifically accumulated at the moment of radicle protrusion. Germination was strongly delayed in the presence of DL-propargylglycine, a specific inhibitor of methionine synthesis. Furthermore, this compound totally inhibited seedling growth. These results therefore validated the proteomics data and established that methionine synthase and AdoMet synthetase are fundamental components controlling metabolism in the transition from a quiescent to a highly active state during seed germination. Interestingly, the level of AdoMet synthetase fell sharply during *M. truncatula* seed filling [2]. A sharp decrease of methionine synthase was also observed during soybean seed filling [3], and a sharp decrease of both methionine synthase and AdoMet synthetase was observed during wheat endosperm development [7]. Therefore, proteomics unraveled a characteristic feature of seed development across plant species and provided the first demonstration of a metabolic control of seed development and germination. This type of control might be a more general feature, as inferred from the observation that relative abundances of metabolic proteins steadily decreased during soybean seed filling, this trend being established for 82 identified proteins [3] (Figure 13.2).

13.7 CONCLUSIONS

The proteomics data described above can be used to elaborate a model of seed development and germination. A first salient feature emerging from these studies is the importance of metabolic control to maintain quiescence in the dry state and, conversely, to allow awakening from that quiescence during germination. This finding seems now well established by studies with different plants (*Arabidopsis*, *M. truncatula*, soybean, and wheat) and with different systems (e.g., somatic and zygotic embryos, whole seeds, organelles). In particular, a well-documented metabolic block concerns the absence of important metabolic enzymes in mature quiescent seeds, such as enzymes involved in methionine metabolism. Owing to the central role of this sulfur amino acid in metabolism, a control exerted at the level of an enzyme such as AdoMet synthetase will have an impact on a myriad of processes, not only at the level of metabolism (protein synthesis, methylations) but also in regulation of development mediated by ethylene and polyamines. A second salient feature put forward by the proteomic data was the general importance of protein modification in seed development and germination, as exemplified by the very large number of proteins that proved to be the specific targets of thioredoxins and of oxidation through carbonylation. The proteomics studies unraveled a new role for protein carbonylation in seed development and germination that has no counterpart in microorganisms and animals. They also revealed a link between ROS leading to protein carbonylation and redox regulatory events catalyzed by thioredoxin in seeds. The results suggest

that a tuning of such protein modifications might exert a dramatic control upon seed development and germination, which will be the object of future research. Finally, the proteomics data documented the distinct role of stored and nascent mRNA pools in germination and seedling growth. In particular, these data showed the possibility to reinduce the accumulation of maturation proteins from the stored mRNAs during early germination.

13.8 FIVE-YEAR VIEWPOINT

Thanks to proteomics, the past five years have seen a tremendous progress in our understanding of several aspects of seed development and germination, both in reference plants and in crops. Indeed, the large amount of data demonstrates the power of descriptive biology to create novel functional insights—for example, metabolic control and redox regulation of seed developmental processes, mechanisms accounting for desiccation tolerance or dormancy control. Concerning the germination process, proteomics proved to be the method of choice for establishing expression profiling since this process mainly relies on stored proteins and from proteins that are synthesized *de novo* from the stored mRNAs. Therefore, this process is not primarily subject to transcriptional control, although some regulation may occur at this level, as in the modulation of germination rate and uniformity. It is anticipated that future proteomics research will address the question of specific tissue expression of seed proteins, to understand better, in a systems biology approach (see Chapter 46), the functioning of a whole seed from the respective roles of its constituents, namely the embryo, the storage tissues, and the seed coat. From an applied perspective, this will provide specific markers of seed quality and vigor that can be used by the seed industry.

REFERENCES

1. Bewley JD and Black M. 1994. *Seeds: Physiology of Development and Germination*, Plenum Press, New York.
2. Gallardo K, et al. 2003. *Plant Physiol* **133**:1–19.
3. Hajdуч M, et al. 2005. *Plant Physiol* **137**:1397–1419.
4. Hajdуч M, et al. 2006. *Plant Physiol* **141**:32–46.
5. Lin S-K, et al. 2005. *Proteomics* **5**:2140–2156.
6. Finnie C, et al. 2002. *Plant Physiol* **129**:1–12.
7. Vensel WH, et al. 2005. *Proteomics* **5**:1594–1611.
8. Balmer Y, et al. 2006. *J Exp Bot* **57**:1591–1602.
9. Méchin V, et al. 2004. *Phytochemistry* **65**:1609–1618.
10. Osborne TB. 1924. *The Vegetable Proteins*, Longmans, Green, London.
11. Gallardo K, et al. 2001. *Plant Physiol* **126**:835–848.
12. Gruis D, Schulze J, and Jung R. 2004. *Plant Cell* **16**:270–290.
13. Mak Y, et al. 2006. *Funct Integr Genomics* **6**:322–337.

14. Gallardo K, et al. 2002. *Plant Physiol* **129**:823–837.
15. Chibani K, et al. 2006. *Plant Physiol* **142**:1493–1510.
16. Østergaard O, et al. 2004. *Proteomics* **4**:2437–2447.
17. Buchanan BB and Balmer Y. 2005. *Ann Rev Plant Biol* **56**:187–220.
18. Job C, et al. 2005. *Plant Physiol* **138**:790–802.
19. Campo S, et al. 2004. *Proteomics* **4**:383–396.
20. Rajjou L, et al. 2006. *Plant Physiol* **41**:910–923.
21. Boudet J, et al. 2006. *Plant Physiol* **140**:1418–1436.
22. Rajjou L, et al. 2004. *Plant Physiol* **134**:1598–1613.
23. Imin N, et al. 2005. *Plant Physiol* **137**:1250–1260.
24. Lippert D, et al. 2005. *Proteomics* **5**:461–473.
25. Gallardo K, et al. 2002. *Physiol Plant* **116**:238–247.

ENDOSPERM AND AMYLOPLAST PROTEOMES OF WHEAT GRAIN

William J. Hurkman, William H. Vensel, Frances M. DuPont,
Susan B. Altenbach, and Bob B. Buchanan

14.1 INTRODUCTION

Recent proteomic studies of the wheat grain have focused on the endosperm tissue due to its importance to productivity and quality [1–3]. The gluten proteins are of major interest in these studies, since they determine the characteristics of extensibility, elasticity, and gas-holding capacity that are unique to wheat flour doughs. The gluten proteins encompass more than 80% of the endosperm protein and consist of gliadins and glutenin subunits. The gliadins are a polymorphic collection of monomeric proteins soluble in 70% alcohol that are separated into α , γ , and ω subgroups. Each subgroup contains many protein species. For example, it has been estimated that there are as many as 40 different α -gliadins and 20 different ω -gliadins in a single wheat variety. The glutenins consist of HMW glutenin subunits (HMW-GS) and LMW glutenin subunits (LMW-GS) linked by interchain S–S bonds to form large insoluble polymers. Most bread wheat cultivars contain three to five different HMW-GS and 15–20 different LMW-GS. A prominent feature of both gliadins and glutenin subunits is the presence of large regions of repetitive sequences consisting primarily

of proline and glutamine residues. Non-gluten proteins encompass smaller percentages of the endosperm protein, but play critical roles in metabolism, controlling grain development, maturation, and desiccation, as well as protecting against both abiotic and biotic stresses [4]. Some endosperm proteins are also responsible for a variety of human allergies as well as an important human food intolerance known as celiac disease [5, 6].

Wheat presents unique challenges for proteomic studies. First of all, the identification of proteins by MS relies on the availability of comprehensive databases of DNA sequences [7]. Bread wheat is a hexaploid with a genome of ~16,000 Mb, which is considerably larger than rice (420 Mb) or *Arabidopsis* (120 Mb), and the genomic sequence is not yet available. However, major sequencing efforts have focused on the expressed portions of the genome and there are currently over 800,000 ESTs from *Triticum* species in NCBI, many of which were generated from endosperm, embryo, or whole-grain cDNA libraries. A second challenge results from the fact that many proteins are encoded by large gene families in wheat [1]. The complexity of the gluten protein gene families, in particular, and the repetitiveness of the sequences make it difficult to distinguish individual proteins. Additionally, the high proportions of proline and glutamine, coupled with low proportions of lysine and arginine in the gluten proteins, make it difficult to cleave these proteins into fragments that can be identified readily by MS. Wheat also has distinct advantages for proteomic studies. Because the major gluten proteins can be separated on the basis of solubility, subfractions of endosperm proteins can be prepared by extraction with different solvents, making it possible to analyze components that encompass a minor percentage of the total endosperm protein [8]. Additionally, the availability of ditelosomic lines has facilitated the mapping of proteins to specific chromosomes and chromosome regions in wheat, something that is difficult in other plant species [9]. During the last five years, proteomic approaches have led to significant progress in the identification and analysis of proteins of wheat grain. In this chapter, we present a brief overview of methods used to fractionate and identify endosperm proteins and then summarize findings from a wide range of wheat endosperm proteome studies.

14.2 FRACTIONATION AND ANALYSIS OF ENDOSPERM PROTEINS

The dynamic range in abundance of proteins in endosperm tissue varies by six or more orders of magnitude and makes the display of the entire endosperm proteome problematic by any single method. Solubility fractionation techniques based on observations made by Osborne [8] provide effective ways to isolate the major classes of endosperm proteins, increasing proteome coverage. The following examples illustrate a few of the many approaches devised for fractionating endosperm proteins for proteomic studies. Hurkman et al. [10] exploited the solubility properties of these proteins by using KCl, methanol, and SDS to prepare fractions enriched in the abundant gliadins and glutenins and the less abundant methanol-soluble and methanol-insoluble albumins and globulins (Figure 14.1). DuPont and Tanaka [11] devised a sequential extraction method using NaI in propanol to separate and quantify gliadins, glutenins, and albumins plus globulins. Other investigators have used a non-ionic detergent, Triton-X114, to separate

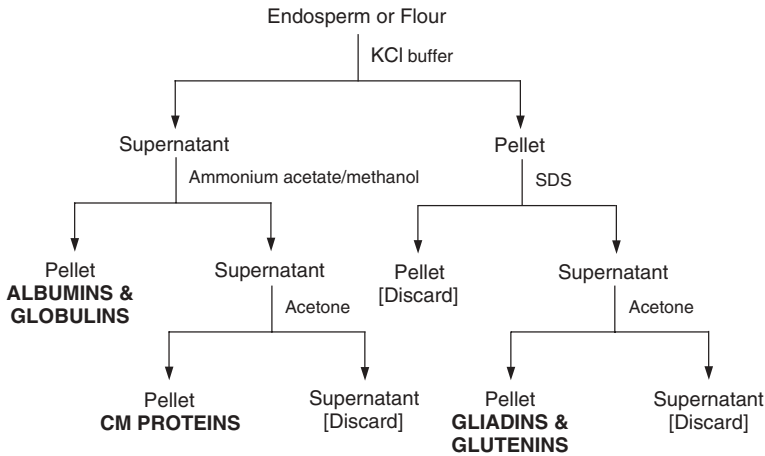


FIGURE 14.1. Sequential extraction scheme for isolation of protein fractions enriched in albumins and globulins, gliadins and glutenins, or CM proteins from wheat endosperm or flour.

amphiphilic endosperm proteins from the salt-soluble proteins by phase partitioning [12, 13].

2-DGE in combination with MS is the method of choice for the global identification of proteins. MALDI-TOF-MS and MS/MS are the two most commonly used methods for the identification of proteins in 2D gel spots. In an analysis of the KCl-soluble albumins and globulins, only about 30% of the proteins could be identified by MALDI-TOF-MS “peptide mass mapping” [14]. Following development of more sensitive nLC-MS/MS techniques for analysis of picogram to nanogram quantities of protein, over 80% of the proteins were identified [4]. In other studies, RP-HPLC was used to obtain large quantities of proteins and protein complexes for detailed sequence analysis by Edman degradation in combination with MS characterization [15, 16]. Prior to analysis by MS, spots are subjected to in-gel digestion with site-specific protease(s), most commonly trypsin, to generate peptide fragments. In contrast to the albumins and globulins, the omega gliadins yield no tryptic fragments and the LMW-GS glutenins and gliadins yield few tryptic peptides amenable to MS analysis. Proteases with other cleavage specificities are needed to fragment the gluten proteins to provide sufficient information for identification. Thermolysin is the most promising of the enzymes tested so far in this connection (Vensel et al., unpublished observations).

The identification of proteins by MS/MS is based on the correlation of MS data with database sequences. Using a “search engine,” fragmentation patterns of enzymatically generated peptides are compared to theoretical data generated from protein sequence databases *in silico* [17]. The ever-increasing rate at which mass spectra can be accumulated means that greater reliance is placed on computer-based software to process spectra and interpret results. Search engines currently available differ in their sensitivity and in their ability to distinguish correct and incorrect peptide hits [18]. The

number of incorrect identifications can be decreased by analyzing all spectra with multiple search engines such as Mascot [19], Sequest [20], and X!Tandem [21] and then validating the combined results using Peptide Prophet [22] and Protein Prophet [23] as implemented in Scaffold (<http://www.proteomesoftware.com/>). Different algorithmic approaches are being developed to improve protein identification searches, because currently available algorithms do not adequately model peptide fragmentation [24].

14.3 THE ENDOSPERM PROTEOME

With its unparalleled resolving power for simultaneously visualizing hundreds of proteins in complex biological samples, 2-DGE has been used extensively to characterize wheat endosperm proteins [25]. Historically, 2-DGE methodology was first developed to separate wheat gliadins in order to study genotypic variation [26, 27]. In the original method, IEF was used for the first dimension and acid PAGE (lactate) for the second dimension [26]. Addition of carrier ampholytes, urea, and non-ionic detergent to the solubilization buffer and the first dimension gel together with SDS and reducing agent to the second dimension gel greatly improved the resolution of 2D gels [28–30]. 2-DGE continues to be utilized to analyze gluten proteins and assign chromosomal location of genes [31–33]. The coupling of 2-DGE with protein MS identification methods made possible more extensive analyses of the entire wheat grain proteome (Table 14.1). In this section, findings are summarized for wheat endosperm proteome studies ranging from characterization of grain development to identification of wheat allergens.

Changes during Grain Development

Knowledge of the metabolic processes taking place during grain development is fundamental to the understanding of both the quality and productivity of grain. Proteomics is providing important information in this area. In an early study, Skylas et al. [34] used 2-DGE and Edman microsequencing to identify 94 proteins in a total wheat endosperm extract. The most abundant proteins were HMW-GS, ω -gliadins, and α -amylase inhibitors and α -amylase/trypsin inhibitors. A preliminary comparison of proteome maps for early (17 DPA) and late (45 DPA) stages of grain fill revealed that PDI and 60S acidic ribosomal proteins were substantially decreased late in development. By preparing a fraction enriched in the albumins and globulins [2] and using a 2-DGE/MS approach, Vensel et al. [4] identified 254 proteins that functioned in 13 metabolic processes ranging from carbohydrate metabolism and protein synthesis/assembly to storage and stress/defense. A comparison of the proteome maps for endosperm protein populations at early (10 DPA) and late stages (36 DPA) of grain fill provided insights into the endosperm developmental program [4]. Carbohydrate metabolism and protein synthesis/assembly were the principal endosperm functions early in grain development (Figure 14.2). Later in development, proteins functional in cell division were absent, but carbohydrate metabolism remained a major function. The predominant activities at this stage were the accumulation of stress/defense

TABLE 14.1. Summary of Wheat Grain Proteomics Studies

Source	Protein Fraction	Number of Proteins Identified	Gluten Proteins	Identification Method	Comments	References
Endosperm	KCl-sol ^a , MeOH-insol ^b	254	2	MALDI ^c , MS/MS ^d	Compared protein profiles in 10- and 36-DPA endosperm	4
Endosperm	KCl-insol	17	17	MALDI, MS/MS	Compared accumulation of gluten proteins under different temperature and fertilizer regimens	42, 43
Endosperm, flour	KCl-sol	68	1	MALDI, MS/MS	Identified thioredoxin target proteins in 10-DPA endosperm	37
Endosperm, flour	KCl-sol, MeOH-sol	57	18	MALDI, MS/MS	Identified thioredoxin target proteins in 40-DPA endosperm	38
Endosperm	KCl-sol	25	1	MALDI, MS/MS	Identified thioredoxin target proteins in 36-DPA endosperm	36
Endosperm, flour	TCA/acetone-insol	94	42	Edman seq	Compared proteins in 17 DPA endosperm and 45-DPA flour	34
Endosperm, flour	TCA/acetone-insol	20	0	MALDI	Identified HSPs at 17 and 45 DPA following heat shock at 15–17 DPA	39
Flour	Urea/thiourea-sol	40	0	MALDI	Compared changes in protein levels between 18/10°C and 34/10°C regimens	40
Flour	TCA/acetone-insol	36	20	Edman seq, MALDI	Compared 6 cultivars for protein differences	46
Flour	Triton X-114-sol	35	0	MALDI	Identified and mapped amphiphilic proteins	12

(continued)

TABLE 14.1. (Continued)

Source	Protein Fraction	Number of Proteins Identified	Number of Gluten Proteins	Identification Method	Comments	References
Flour	Triton X-114-sol	97	0	MALDI	Identified and mapped amphiphilic proteins	48
Dough liquor	Water-sol, NaCl-sol	47	0	MALDI	Identified foam-forming proteins from dough	47
Whole grain	Sodium phosphate buffer-sol	24	0	MALDI	Compared changes in albumins and globulins between 18/10°C and 34/10°C regimens	41
Whole grain	TBS-sol	3	0	MALDI	Identified IgE reactive proteins	52
Whole grain	TBS-sol	3	0	Edman seq	Identified IgE reactive proteins	51
Whole grain	TBS-sol	1	0	Edman seq	Identified IgE reactive proteins	50
Amyloplasts	Urea/thiourea-sol	289	4	MS/MS	Identified proteins in amyloplasts and amyloplast membranes isolated from 10-DPA endosperm	65
Amyloplasts	Urea/thiourea-sol	42	0	MS/MS	Identified thioredoxin target proteins in amyloplasts isolated from 10-DPA endosperm	67
Amyloplasts	Urea/thiourea-sol	156	12	MS/MS	Identified proteins in amyloplasts and amyloplast membranes isolated from 11- to 13-DPA endosperm	63

^a Soluble.^b Insoluble.^c MALDI–TOF–MS.^d ESI–MS/MS.

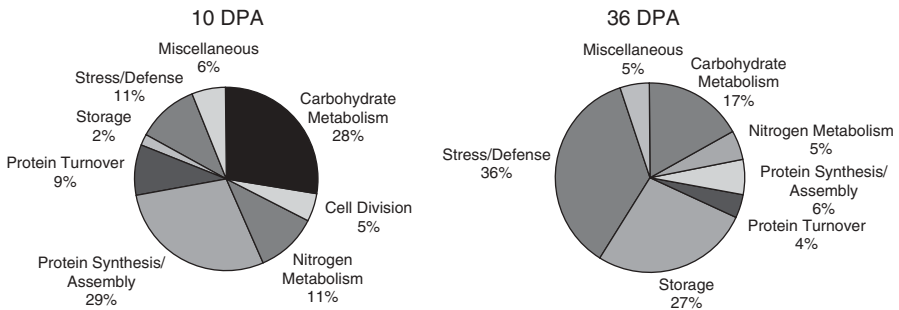


FIGURE 14.2. Functional distribution of proteins recovered in the albumin and globulin fraction of wheat endosperm isolated from developing grain harvested 10 and 36 days post-anthesis. The proteins comprising the functional categories can be found in Vensel et al. [4]. The globulins are the principal storage proteins in the albumin/globulin fraction. The miscellaneous category contains proteins that function in ATP interconversion, lipid metabolism, signal transduction, and transport. See insert for color representation of this figure.

and storage proteins. The proteome maps were also used to identify albumins and globulins targeted by thioredoxin, a widely distributed S–S protein that regulates a growing number of processes [35–37]. Unique sets of targets were found at early (10 DPA) and late (36 DPA) developmental stages [38]. Thus, the proteomic approach has provided new information on both the identification and the regulation of major metabolic pathways functional early and late in grain development.

Effect of Environmental Conditions

Environmental conditions during grain fill alter both wheat yield and flour quality. To elucidate the basis for these changes, comparative proteomics has been used to identify specific protein changes in response to temperature and fertilizer levels during grain fill. Skylas et al. [39] determined the effect of a 3-day high-temperature regimen (40/25°C day/night at 15–17 DPA) on protein profiles in heat-tolerant and heat-sensitive wheat cultivars. At 17 DPA, more proteins responded to high temperature in the endosperm of the heat-tolerant cultivar than to high temperature in the endosperm of the heat-sensitive cultivar. All 17 heat-responsive proteins identified in this study were smHSPs. Because seven proteins responded only in the heat-tolerant cultivar at 17 DPA and one protein remained at high levels at 45 DPA, these proteins were suggested to be candidates for markers for heat-tolerance. Such proteins are of keen interest to plant breeders who are trying to obtain stress-resistant cultivars. In their study, Majoul et al. [40] determined the effect of high temperature (34/10°C day/night) on protein profiles of mature grain. Twenty-five proteins increased in response to high temperature, most of which were thought to function in stress/defense—that is, 70-, 80-, and 83-kD HSPs, chloroplast smHSP, catalase, pectin methylesterase, phosphoinositide-specific PLC, RPM1-like protein, vacuolar ATPase, and LEA protein. The one protein that decreased was glucose-1-phosphate adenylyl-transferase, an enzyme functional in starch biosynthesis. In a second study, Majoul

et al. [41] determined the effect of high temperature on the accumulation levels of albumins and globulins extracted from mature grain. As found in the previous study, the 16 proteins that increased play a role in stress/defense: HSP 82, smHSPs, vacuolar ATP synthase, JIP, and GTP-binding protein. The eight proteins that decreased included mitochondrial ATP synthase and enzymes functional in metabolism—that is, glucose-1-phosphate adenylyltransferase, serine carboxypeptidase, and AdoMet synthetase.

DuPont et al. [42] compared the effect of moderate (24/17°C day/night) and high (37/28°C day/night) temperature regimens with or without post-anthesis fertilizer on gluten protein accumulation during grain development. Relative spot volumes of several α -gliadins, ω -gliadins, and HMW-GS were lower when grain was produced under the moderate temperature regimen in the absence compared to the presence of post-anthesis fertilizer. In contrast, the relative amount of a major LMW-GS was lower in the presence of post-anthesis fertilizer. Compared to the 24/17°C regimen in the absence of post-anthesis fertilizer, relative amounts of some of the α -gliadins and HMW-GS were higher in grain produced under the high-temperature regimen with or without post-anthesis fertilizer. In contrast, the relative amount of a major LMW-GS decreased under the high-temperature regimen. DuPont et al. [43] also demonstrated that the accumulation rates increased more for the sulfur-poor proteins (ω -gliadins, HMW-GS) than for the sulfur-rich proteins (α - and γ -gliadins, LMW-GS) in grain grown under both temperature regimens in the presence of post-anthesis fertilizer or under the high-temperature regimen alone. Thus, proteomics approaches have led to increased understanding of the specific effects of environmental conditions on gluten and non-gluten proteins during grain development.

Cultivar Identification

Wheat grain proteomics is being investigated as an efficient method for identification of cultivars. In a preliminary study, Skylas et al. [44] found significant differences in the protein patterns in the HMW-GS, gliadin, and smHSP regions of 2D gel maps for four cultivars differing in flour quality. These authors concluded that a proteomics approach, especially when applied to the analysis of non-gluten proteins, opens up new possibilities for the development of immunoassays for cultivar identification. In a study comparing two closely related commercial varieties that differed in processing characteristics, Skylas et al. [45] identified 23 proteins that could be used to distinguish the cultivars. Yahata et al. [46] compared the protein maps of seven wheat cultivars and identified 25 proteins, 21 of which were gliadins and glutenins, that could be used as cultivar-specific markers.

Characterization of Protein Fractions

Proteomics has been used to identify proteins in a variety of subfractions of wheat endosperm or flour. Salt et al. [47] identified foam-forming proteins from wheat flour doughs that may play roles in stabilizing gas bubbles in dough and influencing the crumb structure of bread. Of the 42 proteins identified, the most abundant were

α -amylase/trypsin inhibitors, tritin, serpins, and β -amylase. Amiour et al. [12, 48] identified nearly 100 proteins in a fraction containing amphiphilic proteins linked to dough quality. Using multiple-regression analysis, puroindoline a and b, dehydrin, POX, CHS, and RIP were determined to be significantly associated with grain hardness, an important quality attribute. Wong et al. [37] identified proteins in a salt-soluble, methanol-soluble endosperm fraction. The principal proteins in this fraction were α -amylase inhibitors and α -amylase/trypsin inhibitors, proteins historically designated as CM proteins [49], as well as some gliadins. Most of the proteins in this fraction contain conserved cysteines and are potential targets of thioredoxin [36]. Reduction of these proteins by thioredoxin altered their solubility, supporting the view that mobilization of seed reserves is promoted by a thioredoxin-linked signaling mechanism [38].

Allergen Identification

Proteomics offers an effective approach for identifying wheat allergens. 2-DGE was coupled with Edman microsequencing [50, 51] or MS [52] to identify wheat proteins involved in baker's asthma, an IgE-mediated allergy. Proteins that cross-reacted with sera from asthmatic patients include α -amylase inhibitor [50, 51]; GAPDH, TPI, and serpin [52]; and acyl-CoA oxidase and FBA [51]. Wheat food allergies are also IgE-mediated; proteins that cross-react with sera from allergic patients include lipid transfer protein (LTP) as well as α -, β -, γ -, and ω -gliadins [53]. Proteomic procedures are also being developed for the screening of allergens with S–S bonds [54]. Allergen research appears to be an especially promising area for applying proteomics to identify and characterize allergens in foods and food products.

14.4 THE AMYLOPLAST PROTEOME

Amyloplasts are specialized organelles that function in the synthesis and storage of starch, a major component of wheat endosperm. Amyloplasts are related to chloroplasts in that both differentiate from proplastid precursors. Like chloroplasts, amyloplasts are bounded by a double-membrane envelope enclosing an inner compartment, the stroma, which contains soluble proteins and starch granules. In contrast to chloroplasts, amyloplasts are found in non-green tissues and lack complex interior thylakoid membrane structure. As endosperm cells mature, amyloplasts increase in size, fill with starch granules, and become irregular in shape (Figure 14.3). Starch granules are composed mainly of amylose and amylopectin, which are deposited in semicrystalline layers on the exterior of the growing granule by starch synthases and modified by starch branching enzymes [55–57]. Wheat endosperm contains large lenticular-shaped A-type starch granules that form early in development and smaller spherical-shaped B-type granules with lower amyloplast concentrations [58] that develop somewhat later. B-type granules are synthesized in protrusions that interconnect amyloplasts containing A-type starch granules [59, 60]. Similar protrusions, called stromules, are associated with plastids in various organs of *Arabidopsis* and *Nicotiana* [61]. Because chloroplasts

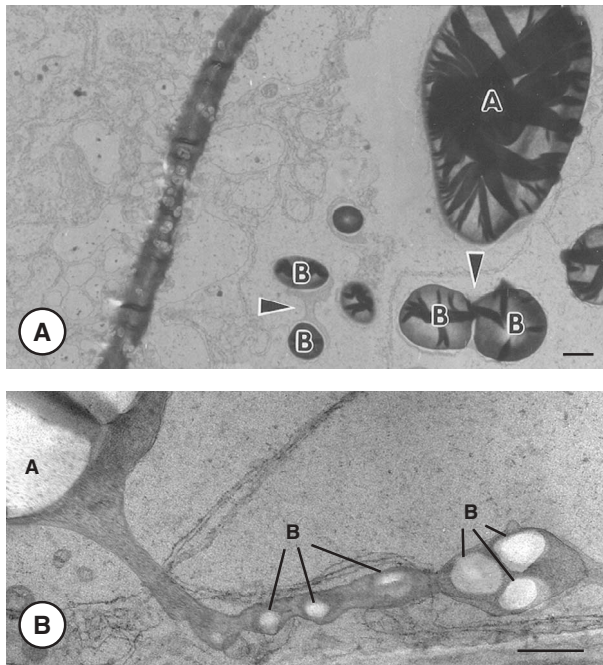


FIGURE 14.3. TEM of wheat endosperm amyloplasts. **A:** Subaleurone cell at 12 DPA containing two B-type starch granules (B) in a stromule (arrows) and an A-type starch granule (A) (Reprinted from Bechtel and Wilson [59]). **B:** Serial section composite of an 11-DPA endosperm cell with an amyloplast containing an A-type granule interconnected with a stromule containing B-type starch granules (Reprinted from Langeveld et al. [60]). Bars represent 1 μm .

are the site of many metabolic processes crucial to the plant cell, Köhler et al. [62] proposed that stromules enhance the functional interactions of chloroplasts with the cytoplasm and other organelles. Bechtel and Wilson [59] suggested that the formation of B-type starch granules in amyloplast protrusions might provide a different environment for starch granule initiation and growth, explaining why they differ from the A-type granules in shape and composition.

Amyloplast Isolation

Although it is well known that chloroplasts play a central role in the metabolism of leaf cells, little is known of amyloplasts aside from their role in starch biosynthesis. Amyloplasts, including those from wheat, are difficult to isolate, because of the fragility of the membrane envelope and the density of the starch granules. Early in grain development, when little starch is present, intact amyloplasts can be isolated by density-gradient centrifugation under hyperosmotic conditions. As grain fill continues, starch granules become much larger in size and isolation of intact amyloplasts becomes problematic. The much denser starch granules break through the fragile amyloplast

envelope during centrifugation, resulting in loss of stromal contents. As with any method of subcellular fractionation, there are questions of purity. Amyloplast preparations may be contaminated with other membrane-bound cellular components such as mitochondria, protein bodies, ER, Golgi apparatus, and peroxisomes. The degree of contamination by other organelles may be evaluated by assays for specific marker enzymes [63–64].

Protein Identification

Andon et al. [63] isolated amyloplasts from grain harvested at 11 and 13 DPA and identified 108 proteins in amyloplasts and 63 in an amyloplast membrane fraction. The majority of the amyloplast proteins functioned in energy metabolism and protein destination/storage. Proteins participating in energy metabolism, ribosomal/transcription regulation, and protein destination/storage were the predominant species in the membrane fraction. The amyloplast preparation probably contained protein bodies or ER since residents of these compartments—gliadins, glutenins and PDI—were among the proteins identified. A large proportion of the proteins were not identified, possibly because the study was undertaken before public databases contained extensive collections of wheat ESTs. More recently, Balmer et al. [65] identified a total of 289 proteins in soluble and membrane fractions from amyloplasts isolated from 10 DPA endosperm. Of these, 119 were found only in the soluble fraction, 83 only in the membrane fraction, and 87 in both fractions. Overlap with proteins identified in the endosperm study of Vensel et al. [4] was relatively small; only 46 of the proteins identified were common to the endosperm and amyloplast studies. Questions of cross-contamination with the amyloplast and amyloplast membrane fractions must be resolved. As discussed by Balmer et al. [65], it is not trivial to assign a protein to a specific compartment. For example, some nuclear proteins that are translocated into multiple compartments (e.g., mitochondria and chloroplasts) may have similar sequence properties, making it difficult to distinguish organelle-specific isoforms. Although Balmer et al., [65] estimated that 90% of the proteins identified were plastid-located, a closer examination of the data suggests that at least 20% of the proteins should be assigned to other cellular compartments [66]. The identification of actin, PDI, GTP-associated proteins, gliadin, and mitochondrial F₀-ATPase in the amyloplast fraction is consistent with the respective presence of components of the cytoskeleton, ER, protein body, Golgi apparatus, and mitochondria. Plastid-specific proteins can be predicted through homology with proteins known to be plastid-located in other tissues or plant species and by the presence of a plastid transit peptide in the N-terminus as predicted by cDNA or genomic sequence information. Although the amyloplast fraction was highly enriched in proteins specific to this organelle, it is clear that membrane vesicles with other subcellular origins co-purified to some extent with the amyloplasts.

Pathway Elucidation

The list of amyloplast proteins identified by Balmer et al. [65] includes the enzymes for a number of complete or nearly complete metabolic pathways. One such pathway for the synthesis of amino acids is illustrated in Figure 14.4. Interestingly, only

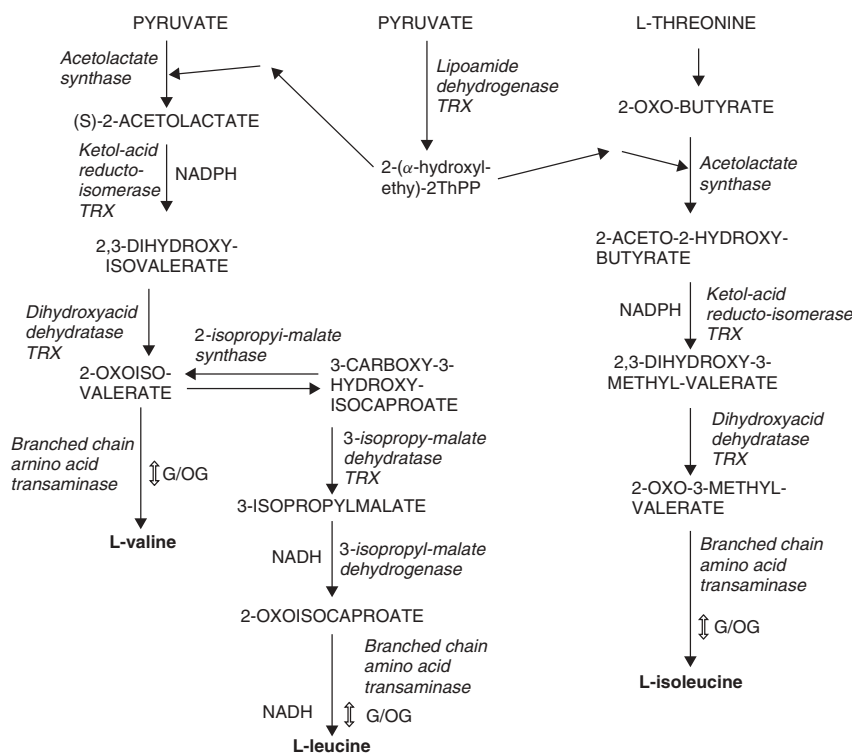


FIGURE 14.4. Enzymes of the branched-chain amino acid pathway identified in the wheat amyloplast proteome preparation: Acetylactate synthase 4.1.3.18; branched-chain amino acid transaminase 2.6.1.42; dihydroxy-acid dehydratase 4.2.1.9; ketol-acid reductoisomerase 1.1.1.86; 3-isopropylmalate dehydratase, large subunit 4.2.1.33; 3-isopropylmalate dehydratase, small subunit 4.2.1.33; 3-isopropylmalate dehydrogenase 1.1.1.85; 2-isopropylmalate synthase 2.3.3.13. Abbreviations: G, glutamate; OG, 2-oxyglutarate; G/OG, conversion of glutamate to 2-oxyglutarate; TRX, enzymes identified as thioredoxin targets.

three of the eight enzymes in Figure 14.4 were also identified in the salt extract of total endosperm, namely, ketol-acid reductoisomerase, 2-isopropyl malate synthase, and lipoamide dehydrogenase (LDH). All of the proteins except ketol-acid reductoisomerase and 2-isopropyl malate synthase are recognized as plastid proteins in the NCBI database and gene sequences for the latter two proteins have N-terminal sequences that are predicted to be plastid transit peptides by the program Target P [67]. Other pathways likely to operate in the wheat endosperm amyloplast include those for glycolysis, starch biosynthesis, pentose phosphate cycle, citric acid cycle, folate 1-carbon metabolism, biosynthesis of most amino acids, nucleotide biosynthesis, fatty acid biosynthesis, and vitamin biosynthesis. This finding indicates that amyloplasts share many of the metabolic functions found in chloroplasts of green tissue [68].

Recent proteomic studies of the amyloplast suggest that it plays a major role in the synthesis of amino acids as well as starch. Within this organelle these are a number of intersections of the pathways for the synthesis of carbohydrates and amino acids appear to reside within the organelle amyloplast. These intersections may represent control points that regulate the balance between starch and protein and thus may be important in understanding the trade-off between grain yield and protein content. For example, the enzyme LDH (Figure 14.4) is essential in pathways converting pyruvate to sugars or amino acids. Furthermore, this enzyme, like others in Figure 14.4, may be regulated by thioredoxin. It is thus clear that recent proteomics studies have provided new evidence for the role of amyloplasts in a spectrum of biosynthetic pathways, a finding in keeping with earlier results obtained with chloroplasts.

14.5 CONCLUSIONS

Advances in proteomics and genomics have improved our understanding of the gluten proteins, a complex and functionally important protein group. Proteomic approaches also have been used to identify other proteins that may play a role in wheat flour functionality, to assign genes for gluten proteins to specific chromosomes, to evaluate protein accumulation profiles during grain development, to investigate redox regulation in the endosperm, to distinguish different wheat varieties, to provide insight into changes in the proteome that occur during grain development as a result of environmental conditions, to identify proteins and metabolic pathways in amyloplasts and to investigate redox regulation in the endosperm. Proteomics research is leading to a better understanding of the manner in which environment impacts the biosynthesis and accumulation of protein and starch in the endosperm and, ultimately, the quality of flour.

14.6 FIVE-YEAR VIEWPOINT

Proteomics efforts have resulted in the cataloging of hundreds of wheat grain proteins. Much of the work thus far has been descriptive, but provides a starting place for addressing important questions related to wheat productivity and quality. A number of technical challenges must be addressed in the coming years. With regard to the gluten proteins, improved methods for enzymatic digestion and MS identification of gliadins and glutenin subunits are required for better understanding the roles that individual proteins play in the formation of the glutenin polymer and, hence, in flour quality. Furthermore, it is essential to improve methods for identifying less abundant proteins, particularly those involved in signaling and regulation to understand molecular processes functional during grain development. Proteomics approaches can also be used to examine PTMs such as phosphorylation, glycosylation, and reduction as well as interactions between proteins and other ligands that are likely to be important

in critical processes in the developing grain. Proteomics studies have paid off handsomely in elucidating enzymes linked to thioredoxin, thereby extending the role of this protein to regulation of a spectrum of cellular processes.

Inevitably, isoforms of proteins are observed in the same or complementary protein fractions. Since isoforms may be targeted to different cellular locations and fulfill different functions, there is a need to examine their expression profiles to learn more about their biological significance. More information is needed on the localization of proteins within the wheat grain. Proteomics analysis of organelles such as amyloplasts provides a list of proteins likely to reside in the organelle. These data must be confirmed by *in situ* localization studies or transgenic approaches using marker-tagged proteins to provide a better picture of organelle function. A continuing challenge is to determine the biological significance of the proteins expressed in a tissue at a defined developmental stage under specific growth conditions. Such studies will lead to a better understanding the effect of environment on wheat quality and productivity. Finally, as the quantity of proteomics data increases for wheat, it is important to develop databases to integrate information generated in different laboratories. Some of this need should be met through the efforts of the European Health Grains Initiative, which promises to provide a more user-friendly interface for studies of all aspects of wheat proteomics.

Disclaimer. The mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by the United States Department of Agriculture and does not imply its approval to the exclusion of other products that may be suitable.

REFERENCES

1. Gianibelli MC, et al. 2001. *Cereal Chem* **78**:635–646.
2. Ikeda TM, et al. 2006. *Theoret Appl Genet* **112**:327–334.
3. Shewry PR and Halford NG. 2002. *J Exp Bot* **53**:947–958.
4. Vensel WH, et al. 2005. *Proteomics* **5**:1594–1611.
5. Breiteneder H and Radauer C. 2004. *J Allergy Clin Immunol* **113**:821–830.
6. Ciccocioppo R, et al. 2005. *Clin Exp Immunol* **140**:408–416.
7. Baginsky S and Gruissem W. 2006. *J Exp Bot* **57**:1485–1491.
8. Osborne TB. 1924. *The vegetable Proteins* Longmans, Green, London.
9. Islam N, et al. 2002. *Proteomics* **2**:1146–1155.
10. Hurkman WJ, et al. 2004. *J Cereal Sci* **40**:295–299.
11. DuPont FM and Tanaka et al. 2005. *J Agric Food Chem* **53**:1575–1584.
12. Amiour N, et al. 2002. *Proteomics* **2**:632–641.
13. Branlard G, et al. 2003. *Proteomics* **3**:168–174.
14. Vensel WH, et al. 2002. *J Biomol Tech* **13**:95–100.
15. DuPont FM et al. 2000. *Cereal Chem* **77**:607–614.
16. DuPont FM, et al. 2004. *Theoret Appl Genet* **108**:1299–1308.

17. Johnson RS, et al. 2005. *Methods* (www.Sciencedirect.com) **35**:223–236.
18. Kapp EA, et al. 2005. *Proteomics* **5**:3475–3490.
19. Perkins DN, et al. 1999. *Electrophoresis* **20**:3551–3567.
20. Eng JK, et al. 1994. *J Am Soc Mass Spectrom* **5**:976–989.
21. Craig R and Beavis RC. 2004. *Bioinformatics* **20**:1466–1467.
22. Keller A, et al. 2003. *Anal Chem* **74**:5383–5392.
23. Nesvizhskii AI, et al. 2003. *Anal Chem* **75**:4646–4658.
24. Wysocki VH, et al. 2005. *Methods* (www.Sciencedirect.com) **35**:211–222.
25. Skylas DJ, et al. 2005. *J Cereal Sci* **41**:165–179.
26. Wrigley, CW. 1968. *Science Tools* **15**:17–23.
27. Wrigley, CW. 1970. *Biochem Genet* **4**:509–516.
28. Klose, J. 1975. *Humangenetik* **26**:231–243.
29. O'Farrell, PH. 1975. *J Biol Chem* **250**:4007–4021.
30. Scheele, GA. 1975. *J Biol Chem* **250**:5375–5385.
31. Dumur J, et al. 2004. *Proteomics* **4**:2685–2695.
32. Islam N, et al. 2003. *Proteomics* **3**:307–316.
33. Islam N, et al. 2003. *Proteomics* **3**:549–557.
34. Skylas DJ, et al. 2000. *J Cereal Chem* **32**:169–188.
35. Buchanan BB and Balmer Y. 2005. *Annu Rev Plant Biol* **56**:87–220.
36. Wong JH, et al. 2003. *FEBS Lett* **547**:151–156.
37. Wong JH, et al. 2004. *Phytochemistry* **65**:1629–1640.
38. Wong JH, et al. 2004. *Plant Cell Physiol* **45**:407–415.
39. Skylas DJ, et al. 2002. *J Cereal Sci* **35**:175–188.
40. Majoul T, et al. 2003. *Proteomics* **3**:175–183.
41. Majoul T, et al. 2004. *Proteomics* **3**:505–513.
42. DuPont FM, et al. 2006. *Eur J Agron* **25**:96–107.
43. DuPont FM, et al. 2006. *J Cereal Sci* **44**:101–112.
44. Skylas DJ, et al. 2001. *Proteomics* **1**:1542–1546.
45. Skylas DJ, 2005. *Aust J Agric Res* **56**:145–155.
46. Yahata E, et al. 2005. *Proteomics* **5**:3942–3953.
47. Salt LJ, et al. 2005. *Proteomics* **5**:1612–1623.
48. Amiour N, et al. 1987. *Oxford Surv Plant Mol Cell Biol* **4**:275–334.
49. Garcia-Olmedo F, et al. 1987. *Theor App Genet* **108**:62–72.
50. Posch A, et al. 1995. *Electrophoresis* **16**:1115–1119.
51. Weiss W, et al. 1997. *Electrophoresis* **18**:826–833.
52. Sander I, et al. 2001. *J Allergy Clin Immunol* **107**:907–913.
53. Battais F, et al. 2005. *J Cereal Sci* **42**:109–117.
54. Yano H. 2003. *Anal Chem* **75**:4682–4685.
55. Baldwin PM. 2001. *Starch/Stärke* **53**:475–503.
56. Buleon A, et al. 1998. *Int J Biol Macromol* **23**:85–112.
57. Smith AM, et al. 1997. *Annu Rev Plant Physiol Plant Mol Biol* **48**:67–87.
58. Peng M, et al. 2000. *Plant Physiol* **124**:265–272.

59. Bechtel DB and Wilson JD. 2003. *Cereal Chem* **80**:175–183.
60. Langeveld SMJ, et al. 2000. *J Exp Bot* **51**:1357–1361.
61. Kwok EY and Hanson MR. 2004. *J Microscopy* **214**:124–137.
62. Kohler RH, et al. 1997. *Science* **276**:2039–2042.
63. Andon NL, et al. 2002. *Proteomics* **2**:1156–1168.
64. Tetlow IJ, et al. 2003. *J Exp Bot* **54**:715–725.
65. Balmer Y, et al. 2006. *J Exp Bot* **57**:1591–1602.
66. Dupont FM. 2008. *BMC Plantbiol* **8**:3.
67. Emanuelsson O, et al. 2000. *J Mol Biol* **300**:1005–1016.
68. Balmer Y, et al. 2006. *Proc Natl Acad Sci USA* **103**:2988–2993.

ROOT PROTEOME

Kuo-Chen Yeh, Chyi-Chuann Chen, and Chuan-Ming Yeh

15.1 INTRODUCTION

Root, like leaf and stem, is one of the major organs in all plants. The development of root is well-organized. The primary root tissues are generated from a group of dividing cells called root apical meristem. The tissues are comprised of epidermis, cortex, endodermis, and pericycle that surrounds the vascular tissues of xylem and phloem. Root hairs emerge from epidermis with a special pattern. Elongation of primary roots and root hairs and the formation of lateral roots are highly regulated (Figure 15.1A). Monocot and dicot plants develop different root patterns (Figure 15.1B). In general, roots are instrumental in the uptake of water and various nutrients. To accomplish these functions, symbiosis with microorganisms may be achieved in the root system of some species. In order to cope with the habitat environments, active molecular mechanisms are involved in the developmental programs to enhance nutrients and water acquisition or to enhance pathogen defenses. Nutrients such as N, P, S, and Fe are known important elements for the signals to control the development of root architecture [1]. Environmental stresses like extreme water conditions, pathogens attack, and even symbiotic processes also affect the development of the root system. Root development

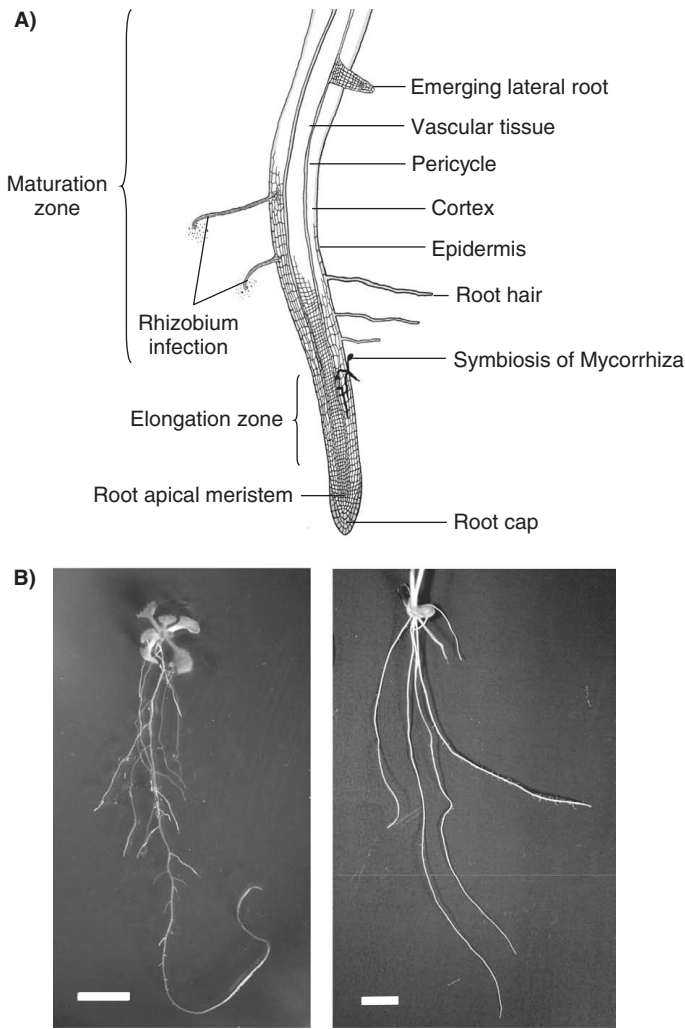


FIGURE 15.1. (A) Scheme of root tissues and their symbiosis with microorganisms. (B) Tap root system and fibrous root system form a dicot plant, *Arabidopsis* (left), and a monocot plant, rice (right). Bars represent 1 cm.

and interaction with environmental stimuli are implicated regulatory networks controlled by a set of genes. Recently, proteomic analysis has become a major approach for the functional characterization of plants. It provides a useful method for identifying proteins associated with a range of root developmental or physiological processes. Proteomics is also used to analyze mutants of root development and for physiological studies analyzing the influences of exogenous stimuli on a particular plant organ and at different developmental stages.

15.2 APPLICATION OF PROTEOMICS IN THE STUDY OF ROOT DEVELOPMENT

To identify genes involved in root development, studies of genetic mutants can be coupled with physiological analysis and proteomic characterization. Several types of lateral and adventitious root mutants have been isolated in cereals [2]. These mutants offer a different background for the study of proteins involved in root developmental processes. A comparison of protein patterns between mutants and wild-type plants revealed genes potentially involved in root differentiation processes or development interactions between lateral roots, primary roots, and adventitious roots [3–5] (Table 15.1).

The mutant *lrt1* from maize (*Zea mays* L.), with defects in the initiation of lateral roots, was used for proteome comparison to search for lateral root development-related genes. Interestingly, about 10% of 150 identified proteins preferentially expressed in *lrt1* seedling roots, including caffeoyl-CoA-3-*O*-methyltransferase, β -glucosidase, L-APX, and 1,4-benzochinone reductase, were involved in lignin metabolism. These up-regulated genes were hypothesized to be repressed by the signals released from lateral roots in the wild-type. Therefore, there may be a signaling pathway between lateral and primary roots involving in the regulation of lateral root initiation. Because *lrt1* mutant lacks lateral roots, the signaling pathway is disrupted and the signals that suppress the transcription of certain genes in the primary root are missing in *lrt1*. In the absence of suppression signals, these genes are up-regulated in the primary roots of *lrt1*. It will be very interesting to find out the target(s) of the signals. During development or when environmental conditions are changed, plants may modify the chemical compositions (e.g., the lignins) of their casparian strip to alter their uptake ability. Because four identified proteins in the *lrt1* mutant are linked to lignin metabolism, the casparian strip may be the target for the signals which may lead to the modification of the casparian strip composition to alter its capacity for water and nutrient uptake [3].

Adventitious root formation is a complex process affected by multiple factors such as hormones, light, or wounding. The mutant *rtcs*, which does not initiate crown roots, a kind of adventitious root in monocot plants, at two early stages of crown root formation in maize, was used for the study of shoot-borne root initiation. Differential protein accumulation during the early stage of coleoptile node development provides information about the molecular processes of crown root initiation, with accumulation at the late stage providing information about their regulatory function during crown root development. Five proteins—histone H2B.2, ARD, GTP-binding protein β subunit, sulfate adenylyltransferase 4, and a putative auxin-binding protein 1 (ABP1)—were identified as being differentially accumulated in the early stage, and 14 were identified in late stage. Interestingly, ABP1 was strongly expressed in 5-day-old coleoptile nodes of wild-type plants but less expressed in 10-day-old ones, whereas its expression in *rtcs* was constitutively strong in 10-day-old plants [4]. The role of ABP1 in the development of crown roots remains to be further characterized. Sorin et al. [5] compared the protein profiles of *sur1-3*, *sur2-1*, *ago1-3*, and *sur2-1/ago1-3 Arabidopsis* mutants defective in the control of adventitious root formation. The regulation of 11 proteins was associated with the number of adventitious root primordia and/or the number of mature adventitious roots [5] (Table 15.1). Interestingly, the expression

TABLE 15.1. List of Root Proteins Regulated by Stimuli

Accession	Identification	Regulation	Reference
NP_012655	F(1)F(0)-ATPase complex, b subunit (<i>Saccharomyces cerevisiae</i>)	AM fungus-induced	18
P49375	F(1)F(0)-ATPase complex, a subunit (<i>Kluyveromyces lactis</i>)	AM fungus-induced	18
MtC00065.1	Acid phosphatase	AM fungus-up-regulated	18
MtC40073	6 Nodulin-like protein	AM fungus-up-regulated	18
MtC10115	Hypothetical protein	AM fungus-up-regulated	18
MtC30193	Aminoacylase-1	AM fungus-down-regulated	18
MtD00070	2-oxoglutarate dehydrogenase	AM fungus-down-regulated	18
MtC10110	Cytosolic phosphoglycerate kinase	AM fungus-down-regulated	18
MtC10055	Lectin	AM fungus-down-regulated	18
MtC30069	Proteasome subunit	AM fungus-down-regulated	18
MtC00127	Ferritin	AM fungus-down-regulated	18
MtC10318	ATP synthase D chain, mitochondrial	AM fungus-down-regulated	18
MtC10146.1	Lipoxygenase	AM fungus-down-regulated	18
MtC00365.1	Hypothetical protein	AM fungus-down-regulated	18
MtC10242	Nucleoside diphosphate kinase	AM fungus-down-regulated	18
MtC00779	Thioredoxin H-type	AM fungus-down-regulated	18
AJ132892	Mthl	<i>G. intraradices</i> -induced	17
MtC30378	MtBcp1	<i>G. intraradices</i> -induced	17
MtBC45C08	Peroxidase 1A precursor (<i>M. sativa</i>)	<i>G. mosseae</i> -induced	15
MtBA40B11	Serine hydroxymethyltransferase (<i>A. thaliana</i>)	<i>G. mosseae</i> -induced	15
MtBC43G02	Glutathione transferase (<i>G. max</i>)	<i>G. mosseae</i> -induced	15
MtBA50B09	Cytochrome- c-oxidase subunit 6b (<i>A. thaliana</i>)	<i>G. mosseae</i> -induced	15
MtBC41E10	Myosin heavy chain-like protein (<i>A. thaliana</i>)	<i>G. mosseae</i> -induced	15
MtBC43C07	Profucosidase precursor (<i>P. sativum</i>)	<i>G. mosseae</i> -induced	15
MtBB50E11	Enolase (<i>Lycopersicon esculentum</i>)	<i>S. meliloti</i> -induced	15
sp P27993	Leghemoglobin 2 (<i>M. truncatula</i>)	<i>S. meliloti</i> -induced	15
MtBC21G08	Polygalacturonase inhibitor protein (<i>Pyrus communis</i>)	<i>G. mosseae</i> up-regulated	15
AAB63603	Triosephosphate isomerase	Salt-up-regulated	19

TABLE 15.1. (Continued)

Accession	Identification	Regulation	Reference
BAB64646	Putative splicing factor-like protein	Salt-up-regulated	19
BAA76393	Cytochrome c oxidase subunit 6b-1	Salt-up-regulated	19
AAC49818	Peroxidase	Salt-induced	19
AAO65861	Putative actin-binding protein	Salt-up-regulated	19
Q42971	Enolase	Salt-down-regulated	19
BAB69069	UDP-glucose pyrophosphorylase	Salt-down-regulated	19
P14654	Glutamine synthetase (GS) root isozyme	Salt-down-regulated	19
P93438	S-adenosylmethionine synthetase 2	Salt-down-regulated	19
BAB89723	Putative nascent polypeptide-associated complex alpha chain	Salt-down-regulated	19
P23346	Superoxide dismutase 5 [Cu– Zn] 4AP	As-up-regulated	29
Q8L8G3	Glutathione peroxidase	As-up-regulated	29
Q8LK64	Putative glutathione peroxidase	As-up-regulated	29
6626536	Peroxiredoxin	As-up-regulated	29
19437269	Putative 1,4 benzoquinone reductase	As-up-regulated	29
9731241	ATP synthase	As-up-regulated	29
B83446	Succinyl-CoA synthetase	As-up-regulated	29
15196515	Cytochrome P450	As-up-regulated	29
P49027	Guanine nucleotide-binding protein beta subunit	As-up-regulated	29
P23345	Superoxide dismutase [Cu– Zn] 4	As-down-regulated	29
At2g26560	Similar to latex allergen (<i>Hevea brasiliensis</i>), patatin-like phospholipase	Cd-induced	24
At1g13440	Glyceraldehyde 3-phosphate dehydrogenase	Cd-induced	24
At5g59190	Peptidase/subtilase	Cd-induced	24
At3g22890	ATP sulfurylase, APS1	Cd-up-regulated	24
At4g02520	Atpm24.1 glutathione S-transferase, AtGSF2	Cd-up-regulated	24
At1g02920	Glutathione S-transferase, AtGSTF7	Cd-up-regulated	24
At1g02930	Glutathione S-transferase, AtGSTF6	Cd-up-regulated	24
At4g23670	Putative major latex protein, Bet v I, PR-10 family	Cd-up-regulated	24
At5g65140	Trehalose-6-phosphate phosphatase	Cd-up-regulated	24

(continued)

TABLE 15.1. (Continued)

Accession	Identification	Regulation	Reference
At1g36440	Hypothetical protein	Cd-up-regulated	24
At1g17290	Alanine transaminase (ALAAT1)	Cd-up-regulated	24
At5g64250	2-Nitropropane dioxygenase-like protein	Cd-up-regulated	24
At3g48670	Putative protein; MIPS: myosin heavy chain	Cd-up-regulated	24
At5g11670	NADP-dependent malic enzyme	Cd-up-regulated	24
At1g36370	Glycine hydroxymethyltransferase	Cd-up-regulated or -induced	24
At5g57410	Unknown protein	Cd-down-regulated	24
At5g11170	DEAD box RNA helicase RH15-like	Cd-down-regulated	24
MtGI TC 39287	Disease-resistance response protein pi 49	<i>A. euteiches</i> -induced	32
MtGI TC 28623	18.2-kDa class I heat shock protein	<i>A. euteiches</i> -induced	32
MtGI TC 31868	ABA-responsive protein ABR17	<i>A. euteiches</i> -up-regulated	32
MtGI TC 31872	ABA-responsive protein ABR17	<i>A. euteiches</i> -up-regulated	32
MtGI TC 31872	ABA-responsive protein ABR17	<i>A. euteiches</i> -up-regulated	32
MtGI TC 29943	Pathogenesis-related protein class 10 [PR10]	<i>A. euteiches</i> -up-regulated	32
MtGI TC 36563	18.2-kDa class I heat shock protein	<i>A. euteiches</i> -up-regulated	32
MtGI TC 31880	Proline-rich protein	<i>A. euteiches</i> -up-regulated	32
MtGI EST 10405	Partially similar to glycine-rich cell wall structural protein	<i>A. euteiches</i> -up-regulated	32
MtGI TC 33591	Isoliquiritigenin 2- <i>O</i> -methyltransferase	<i>A. euteiches</i> -up-regulated	32
MtGI TC39644	Cold acclimation-specific protein CAS18/dehydrin-like protein	<i>A. euteiches</i> down-regulated	32
L11601	Glutathione <i>S</i> -transferase	Ethylene-up-regulated	7
NP_196527	Inorganic pyrophosphatase-like protein	Ethylene-up-regulated	7
15235401	Glutathione <i>S</i> -transferase PM24 (24-kDa auxin-binding protein)	Gravity-up-regulated	14
15220216	Ca ²⁺ -dependent membrane-binding protein annexin	Gravity-up-regulated	14
4587541	Lipase/acyl hydrolase with GDSL-motif family	Gravity-up-regulated	14

TABLE 15.1. (Continued)

Accession	Identification	Regulation	Reference
18404382	Malate dehydrogenase	Gravity-up-regulated	14
15231715	Fructose biphosphate aldolase-like protein	Gravity-up-regulated	14
18390831	Elongation factor 1 α	Gravity-up-regulated	14
15242516	Actin 2/7 (ACT2/7)	Gravity-down-regulated	14
15233627	Tubulin α 6 chain (TUA6)	Gravity-down-regulated	14
15240054	Alcohol dehydrogenase class III	Gravity-down-regulated	14
15239772	Aspartate aminotransferase, cytoplasmic isozyme 1	Gravity-down-regulated	14
18415909	ATP synthase β chain 1	Gravity-down-regulated	14
15236375	Glycine hydroxymethyltransferase	Gravity-down-regulated	14
15218869	Isocitrate dehydrogenase	Gravity-down-regulated	14
15239571	Quercetin 3-O-methyltransferase 1	Gravity-down-regulated	14
15238686	Vitamin B ₁₂ -independent methionine synthase	Gravity-down-regulated	14
15241847	Heat shock cognate 70-kDa protein 2 (HSP70-2)	Gravity-down-regulated	14
BAB92583.1	Putative 1,4-benzoquinone reductase (<i>O. sativa</i>)	Up-regulated in <i>lrt1</i> mutant	3
Q99069	Glycine-rich RNA-binding protein (<i>Z. mays</i>)	Up-regulated in <i>lrt1</i> mutant	3
AAL08496	L-Ascorbate peroxidase (<i>H. vulgare</i>)	Up-regulated in <i>lrt1</i> mutant	3
AF404770	40 S ribosomal protein S15 (<i>E. oleifera</i>)	Up-regulated in <i>lrt1</i> mutant	3
BAA78733	Caffeoyl-CoA 3-O-methyltransferase (<i>O. sativa</i>)	Up-regulated in <i>lrt1</i> mutant	3
1E4LA	β -Glucosidase (<i>Z. mays</i>)	Up-regulated in <i>lrt1</i> mutant	3
U89341	Phosphoglucumutase1 (<i>Z. mays</i>)	Up-regulated in <i>lrt1</i> mutant	3
L39014	Protein disulfide isomerase (<i>Z. mays</i>)	Up-regulated in <i>lrt1</i> mutant	3
P30756	Histone H2B.2	Down-regulated in <i>lrt1</i> mutant	4
T02918	ARD	Down-regulated in <i>lrt1</i> mutant	4
P25387	GTP-binding protein b subunit (<i>C. reinhardtii</i>)	Induced in <i>rtcs</i> mutant	4
AAB64207	Elongation factor 1 α (<i>Z. mays</i>)	Induced in <i>rtcs</i> mutant	4
AAA72022	Mn-superoxide dismutase (<i>Z. mays</i>)	Induced in <i>rtcs</i> mutant	4
CAA64683	Osr40cl (<i>O. sativa</i>)	Induced in <i>rtcs</i> mutant	4
XP_470282	Embryo-specific protein (<i>O. sativa</i>)	Induced in <i>rtcs</i> mutant	4

(continued)

TABLE 15.1. (Continued)

Accession	Identification	Regulation	Reference
BAD32975	Ubiquitin-conjugating enzyme (<i>O. sativa</i>)	Induced in <i>rtcs</i> mutant	4
AAM63309	Sulfate adenylyl-transferase 4	Up-regulated in <i>rtcs</i> mutant	4
XP_476338	Ethylene-inducible protein (<i>O. sativa</i>)	Up-regulated in <i>rtcs</i> mutant	4
AAX09965	Protein disulfide isomerase	Up-regulated in <i>rtcs</i> mutant	4
AAB04687	Histone H2A (<i>Z. mays</i>)	Up-regulated in <i>rtcs</i> mutant	4
CAA40061	Auxin-binding protein (<i>Z. mays</i>)	Up-regulated in <i>rtcs</i> mutant	4
AAC27702	Cytosolic 6-phosphogluconate dehydrogenase (<i>Z. mays</i>)	Up-regulated in <i>rtcs</i> mutant	4
AAU93346	Actin (<i>S. officinarum</i>)	Down-regulated in <i>rtcs</i> mutant	4
AAB86960	Profilin (<i>Z. mays</i>)	Down-regulated in <i>rtcs</i> mutant	4
AAB81110	Triosephosphate isomerase 1 (<i>Z. mays</i>)	Down-regulated in <i>rtcs</i> mutant	4
P27164	Calmodulin-related protein	Down-regulated in <i>rtcs</i> mutant	4
AAN31855	SAM synthetase (<i>A. thaliana</i>)	Down-regulated in <i>rtcs</i> mutant	4
AF045770	Methylmalonate-semialdehyde dehydrogenase	Auxin and zinc-induced	8
AY035011	Elongation factor	Auxin and zinc-induced	8
AC079685	NADPH-dependent oxidoreductase	Auxin and zinc-induced	8
P29545	Elongation factor 1 β'	Auxin and zinc-induced	8
AP003269	Nascent polypeptide associated complex α chain	Auxin and zinc-induced	8
Q05431	L-Ascorbate peroxidase	Auxin and zinc-induced	8
AF046884	Late embryogenesis abundant protein	Auxin and zinc-induced	8
D50301	Aldolase C-1	BL-up-regulated	10
AB003328	MADS box-like protein	BL-up-regulated	10
B028603	Alpha-subunit of GTP-binding protein	BL-up-regulated	10
P12653	Glutathione <i>S</i> -transferase	BL-up-regulated	10
Z48728	Type-1 pathogenesis-related protein	BL-up-regulated	10
L18914	Calmodulin	BL-down-regulated	10
AF032703	NBS-LRR-type resistance protein	BL-down-regulated	10
Y08987	OSR40 protein	BL-down-regulated	10
P33044	Antifungal protein R	BL-down-regulated	10
P28756	Superoxide dismutase [Cu–Zn] 1	BL-down-regulated	10
P28757	Superoxide dismutase [Cu–Zn] 2	BL-down-regulated	10

TABLE 15.1. (Continued)

Accession	Identification	Regulation	Reference
19387274	Putative receptor-like protein kinase	<i>Azoarcus sp.</i> - and JA-induced	11
46399155	Membrane-binding protein (putative receptor-like protein kinase)	<i>Azoarcus sp.</i> - and JA-induced	11
33440014	Pathogenesis-related protein (Prb1)	<i>Azoarcus sp.</i> - and JA-induced	11
134190	Salt-stress-induced protein (SalT)	<i>Azoarcus sp.</i> - and JA-induced	11
9230757	Pathogenesis-related protein (PR-10b)	<i>Azoarcus sp.</i> - and JA-induced	11
38678114	Pathogenesis-related protein (RSOsPR10)	<i>Azoarcus sp.</i> - and JA-induced	11
7619799	Bowman–Birk trypsin inhibitor, putative	JA-induced	11
5360230	GTP-binding nuclear protein Ran	JA-induced	11
5852087	Zwh0010.1 (oxalate oxidase/germinlike protein)	JA-induced	11
22830909	Glucan <i>endo</i> -1,3-beta-D-glucosidase, putative	JA-induced	11
476550	Trypsin inhibitor (Bowman–Birk)	JA-induced	11
7442204	Probenazole-induced protein PBZ1 (PR-10a)	JA-induced	11
51536083	1-Aminocyclopropane-1-carboxylate oxidase, putative (ACC oxidase)	JA-induced	11
34914740	Glutathione <i>S</i> -transferase, putative	JA-induced	11
5777629	Class III peroxidase 59 precursor	JA-induced	11

of three auxin-inducible GH3-like proteins—GH3-3, AtGH3a, and DFL1—was positively correlated with the number of mature adventitious roots. This observation was consistent with the positive relation between endogenous auxin production and number of adventitious roots [5]. These auxin-inducible proteins may be involved in the mechanism of adventitious root formation.

In addition to mutants of later root initiation and adventitious root formation, 24 *Arabidopsis* genes are involved in root hair development [6]. These mutants are distributed in different root hair development processes, including root hair initiation,

root hair bulge, and tip growth. We are looking forward to the study of these mutants to understand proteins involved in root hair formation.

15.3 PROTEOME ANALYSIS OF RESPONSE TO PLANT HORMONES

Ethylene has been shown to play an important role in root hair development in *Arabidopsis*. Root hair formation in young *Arabidopsis* roots is inhibited and enhanced by an effective inhibitor of 1-aminocyclopropane-1-carboxylate (ACC) synthase, aminooxyacetic acid (AOA) and ACC treatments, respectively. *Arabidopsis* seedlings pretreated with either 250 μ M AOA or a control medium were incubated with or without 100 μ M ACC followed by monitoring the proteome of root hair formation at various time points. Three distinct GST isoforms—AtGSTF2, AtGSTF8, and AtGSTU19—expressing during the formation of root epidermis were identified. Among these isoforms, AtGSTF2 protein was specifically up-regulated by ethylene [7] (Table 15.1). Its steady-state mRNA level was greatly reduced in the roots of the ethylene-insensitive mutant *ctr1*. The ethylene-induced AtGSTF2 might play a specific role in the process of early root development of *Arabidopsis* seedlings.

Auxin and Zinc (Zn) regulate callus and root formation in rice. Zn is involved in the biosynthesis of IAA from tryptophan and suggested to play multiple roles in auxin-induced developmental processes [8]. In the rice suspension culture system, seven proteins—methylmalonate-semialdehyde dehydrogenase (MMSDH), elongation factor, NADH-dependent oxidoreductase, EF-1 β' , nascent polypeptide associated complex α chain, L-APX, and LEA protein—were identified to be induced by 2,4-D and Zn [8] (Table 15.1). Interestingly, the expression of MMSDH mRNA was also induced by exogenous GA₃ treatment in wild-type and was higher in a constitutive GA response mutant *slr* of rice than in wild-type [9]. According to these results, MMSDH may play a role in root development stimulated by auxin and/or Zn and GA in rice.

The application of brassinolide (BL), a plant steroid hormone, helps to increase crop yields, possibly through its function in conveying the message for environmental stress tolerance and disease resistance. Proteins related to stress tolerance were found in the root treated with BL [10]. A total of 508 proteins were detected in the rice root tissue on 2-DGE, and the expression of 12 proteins changed on BL treatment (Table 15.1). BL-induced proteins such as SOD, antifungal protein R, and PR-1 may play roles in the defense against stresses.

JA is a plant hormone involved in pathogen defense. In the root of both endophyte-compatible and less compatible rice cultivars, JA can strongly induce the expression of similar proteins, including PR-10a. These JA-induced proteins in rice are Bowman–Birk trypsin inhibitors, germin-like protein, putative endo-1,3- β -D-glucosidase, GST, 1-propane-1-carboxylate oxidase synthase, POX precursor, PR-10a, and a RAN protein previously not found to be JA-induced [11] (Table 15.1).

From a comprehensive survey of these studies, it is interesting that the expression of GST proteins is affected by several hormones. GSTs with high affinity for auxins [12] and cytokinin [13] have also been suggested. They may participate in hormone homeostasis or induced responses in plants and may play a role in root development and stress tolerance.

15.4 PROTEOME ANALYSIS OF GRAVITY RESPONSE

The gravity-stimulated proteome of *Arabidopsis* roots was also recently surveyed [14]. Gravity regulated the expression of proteins in *Arabidopsis* root apices. Ten proteins showed decreased expression on gravitational stimulation, and six proteins exhibited increased expression (Table 15.1). These proteins are involved in the structure of the cytoskeleton and the calcium signal transmission system. Especially, the β subunit of the E1 component of PDH, FBA, and 29S proteasome β subunit E1 protein appeared in two different MW types during gravitational stimulation. The molecular-weight shift change of the three proteins to the gravitropism was first uncovered in plants.

15.5 ACTION TO MICROBE-PLANT SYMBIOSIS

To acquire nutrients and water, lateral roots and root hairs develop to increase root surface. These derivatives also have contact with microorganisms. Two major kinds of microbial symbiosis are beneficial to plants through the interaction with roots. AM fungi supply plants with phosphate and other nutrients. The interaction of plant and fungi is nonspecific. About 80% of land plants are mycorrhized by fungi belonging to the order *Glomales*. By contrast, legume plants and their symbionts, nitrogen-fixing bacteria, have highly specific partner pairs. Host specificity is determined by the signal exchanges between legume plants and *Rhizobium* bacteria. Plants secrete flavonoid or isoflavonoid compounds from their root system to induce the biosynthesis of Nod factors, lipo-chitin-oligosaccharides, in the counterpart *Rhizobium* bacteria. The Nod factors act as morphogens to the root hairs or root system of the host to ready the plant for bacterial invasion. The successful perception of this host-strain-specific interaction results in the development of symbiosis nitrogen-fixing nodules.

M. truncatula is the model legume for the study of both kinds of symbiosis. The genome project has been formed by an international collaboration, with websites for *M. truncatula*-related resources at <http://www.medicago.org/>. The symbiosis systems between *M. truncatula* and the rhizobium *S. meliloti* and the AM fungi *G. intraradices* or *G. mosseae* are now models for proteomics study. Symbiosis-related proteins from *M. truncatula* were surveyed in the earlier years of the century, when sequence information about *M. truncatula* was scarce. Several proteins specifically related to AM fungus or *S. meliloti* symbiosis were identified (Table 15.1). The Dumas-Gaudot group has invested effort into establishing the identities of membrane-associated proteins by 2-DGE, LC-MS/MS, and 1D LC-MS/MS for further study of the symbioses [16, 17]. The membrane proteins associated with AM fungus interaction were initially identified in *M. truncatula* [17, 18]. In addition to previously identified genes, acid phosphatase, lectin, and several novel proteins regulated by the action of AM fungus symbiosis were revealed in these studies (Table 15.1). Two proteins, an efflux H^+ -ATPase (Mtha1) and a blue copper-binding protein (MtBcp1), were specifically detected in only the periarbuscular PM fraction of *G. intraradices*-inoculated roots. These results confirm the methods that can be used to isolate the genes relevant to symbiosis and also reveal the specific molecular markers in the periarbuscular membrane.

Species specificity is an important characteristic of symbiotic interaction. Subtle differences may disrupt the perfect interaction for the symbiosis. Complicated mechanisms, such as ligand–receptor interaction or hormone action, may be involved in the rejection system. For example, external application of JA inhibits endophytic colonization on the rice root. By comparing the proteome of bacterial challenge on JA treatment, Miche et al. [11] found that defense proteins, including PR proteins (Prb1, RSOsPR10), or RLKs, were induced by both JA and endophytic bacterium in the less compatible rice cultivar, whereas only one protein, SalT, with two isoforms, was induced in the susceptible cultivar. JA was suggested to play a role in restricting endophytic colonization in rice, or the endophytic bacterium was not able to induce JA response in the susceptible cultivar. Indeed, elevated levels of JA can prevent successful colonization [11].

15.6 PROTEOME TO COPE WITH ENVIRONMENTAL STRESSES

Environmental stresses, including biotic and abiotic stresses such as pathogen infection, salinity, drought, herbicide and heavy metals, are important factors that affect growth and metabolism of plants. To cope with environment stresses, plants can change their gene expression and protein accumulation [19]. One of the most common consequences of stresses is ample production of ROS and, eventually, oxidative stress [20]. To counteract the harmful effects of oxidative stress, plant cells develop an antioxidant defense system for redox balance [21]. Information from proteomics analysis of plant roots in response to biotic and abiotic stresses reveals many proteins involved in ROS metabolism. A set of antioxidant enzymes was identified under different stress conditions. For example, up-regulation of POX was induced by salt stress; three SODs, two GPX, one peroxiredoxin, and one *p*-benzoquinone reductase were up-regulated by As treatment; three GSTs were abundant under Cd stress; and four proteins—dihydrolipoamide dehydrogenase (LADH), a putative NADP-dependent oxidoreductase, GST, and mannose-6-phosphate reductase—were induced after short-term exposure to Ni. These results are also consistent with data obtained from microarray analysis of response to environmental stresses such as salt, drought, and heavy metals [21–25]. Besides inducing the common stress-inducible category, different stress conditions also induce specific proteomes (Table 15.1). Individual stress proteomes are discussed as follows.

Salinity is one of the most significant limiting factors in agriculture worldwide. About 20% of the land mass and approximately half of all irrigated land is affected by salinity [19]. Therefore, screening salt-stress-responsive genes or proteins will help improve salt tolerance and traits of crops by genetic engineering. Salt-stress-responsive genes and proteins have been identified by cDNA microarray and proteomic analysis in the model plants *Arabidopsis* and rice [23, 25, 26]. Most genes or proteins identified are involved in the resistance mechanisms of ion transport, ROS scavenging, and osmolyte synthesis. Recently, 10 salt stress-responsive proteins, including four previously identified and six novel proteins, were identified in the rice root by 2-DGE and MS analysis (Table 15.1). Interestingly, among six novel proteins, a putative splicing factor-like protein was highly induced by salt stress, which suggests

that pre-mRNA splicing is involved in the salt stress response [19]. Splicing process plays an important role in the regulation of gene expression and contributes greatly to the proteomic complexity in eukaryotes [27]. Forment et al. [28] expressed two *Arabidopsis* SR-like splicing related proteins, AtRCY1 or AtSRL1, in yeast, which induced increased tolerance to salt stress. In addition, overexpression of *AtSRL1* provides salt tolerance to transgenic *Arabidopsis* plants. Although further work is needed to elucidate its accurate function, AtSRL1 protein may play an important role in salt tolerance. In contrast, a putative α -NAC (α -nascent polypeptide-associated complex) was down-regulated by salt stress. NAC, comprised of α and β chains, can reversibly bind to eukaryotic ribosomes and is probably the first cytosolic protein to contact nascent polypeptide chains emerging from the ribosome. Although α -NAC has been suggested to be involved in Alzheimer's disease and Down's syndrome in humans, plant NAC has not been reported. Down-regulation of α -NAC by salt stress might affect the process of transcription and translation.

Heavy metal is a major environment pollutant that can induce severe stress reactions in organisms, with potentially dangerous bioaccumulation through the food chain. Rapid industrialization has accumulated a great deal of toxic metals in the environment. For insight into plant responses to heavy metal toxicity, a proteomic analysis of maize root in response to arsenic toxicity involved maize seedlings treated with sodium arsenate or sodium arsenite for 24 h, with differentially expressed proteins identified by 2-DGE and MALDI-TOF-MS analysis [29]. The proteins highly responsive to As included ROS scavenging enzymes (three SODs, two GPX, one peroxiredoxin, and one *p*-benzoquinone reductase), key enzymes from primary aerobic metabolism (succinyl-CoA synthetase and ATP-synthase), one detoxifying enzyme, cytochrome P450, and one probable signaling protein, guanine nucleotide-binding protein β subunit. More recently, proteome analysis in response to Cd stress in *A. thaliana* roots identified four class proteins including metabolic enzymes (e.g., ATP sulfurylase, glycine hydroxymethyltransferase, and trehalose-6-phosphate phosphatase), GSTs, latex allergen-like proteins, and unknown proteins [24]. The data suggested a possible consequence of activated phytochelatin synthesis by Cd with up-regulated ATP sulfurylase, which catalyses the first step in sulfur assimilation. The result was confirmed by Western blot and microarray analysis [24, 30]. Some heavy metals such as Cu or Zn play roles on the enzyme activity as co-factors. Cu- and/or Zn-binding proteins in *Arabidopsis* roots were isolated by IMAC and identified with proteomics. The putative metal-binding motifs were predicted [31]. The association of Cu or Zn and these proteins implies that metals may participate in their structural conduction and possibly regulate the activity.

A. euteiches, a fungal pathogen, causes the biotic stress, a root rot disease, in several legumes such as pea, alfalfa, and bean and therefore is a major yield-reducing factor in legume crop production. In one study, *M. truncatula* seedlings were inoculated with *A. euteiches* and harvested at several time points between 6 h to 3 weeks after inoculation. Total proteins were extracted and separated by 2-DGE, and differentially expressed protein spots were analyzed by MALDI-TOF-MS. In this study, two putative cell wall proteins, two smHSPs, one isoliquiritigenin 2-*O*-methyltransferase, and one protein with similarities to the cold acclimation specific protein CAS18

were identified. The remaining six proteins differentially expressed in response to the pathogen show similarities to disease-resistance response protein pi 49, PR10, or ABA-responsive protein ABR17 (Table 15.1). Although these six proteins were annotated as different proteins, they show high sequence similarity to each other and all belong to the PR10-like family [32]. Transcriptional profiling of *M. truncatula* roots in response to *A. euteiches* showed that a range of genes encoding PR10- or ABA-responsive proteins was also strongly induced at the transcriptional level [33]. The transcriptomics results are consistent with the proteome data presented above. Some PR10 family proteins are induced by exogenous ABA [34]. In addition, the six PR10-like family proteins show high sequence similarity to previously described ABA-responsive proteins of other plants. These results suggest that ABA may be a general stress response signal in pathogen-infected root tissues and that increased ABA may be involved in the regulation of these PR10-like proteins.

15.7 CONCLUSIONS AND FIVE-YEAR VIEWPOINT

Root development and response involve complicated gene expression and protein accumulation. Proteome analyses, a simple strategy to compare the protein profiles, is expected to reveal differential protein accumulations at different developmental stages or responses to stimuli. However, the information gained from this method is usually limited and discontinuous in terms of physiological significance. Another strategy can be considered to compare responses in mutants that have been anticipated in the up- or downstream relation of root development or responses to stimuli. By the growth of sequence information, technical improvement of proteomics should provide more detail and accurate information to the protein identification in the future. On the other hand, it should be noted that the platforms of proteomics such as the size of 2D gels or type of LC systems may directly affect the resolution and consistency of protein profiling. For the isolation membrane fraction and subcellular organelles, the possibility of contamination should be considered and cautiously prevented. Because hydrophobic proteins are not completely compatible to IEF system, appropriate LC systems may be applied for high-quality membrane protein profiling. To assign biological function of identified proteins with accumulation difference, more experimental data with biochemical and genetic approaches followed by coordination with bioinformatic analyses will be required. The application of high-resolution proteomics methods and coordination of genomics data may help to elucidate the mechanisms involved.

ACKNOWLEDGMENT

We thank Ying Wang for the diagram in Figure 15.1.

REFERENCES

1. Lopez-Bucio J, Cruz-Ramirez A and Herrera-Estrella L. 2003. *Curr Opin Plant Biol* **6**:280–287.
2. Hochholdinger F, et al. 2004. *Trends Plant Sci* **9**:42–48.
3. Hochholdinger F, Guo L, and Schnable PS. 2004. *Plant Mol Biol* **56**:397–412.
4. Sauer M, et al. 2006. *Proteomics* **6**:2530–2541.
5. Sorin C, et al. 2006. *Plant Physiol* **140**:349–364.
6. Cavell A and Grierson CS. 2000. Genetics of root hair development. In: *Root Hairs: Cell and Molecular Biology* (Ridge RW and Emons AMC, eds.), Springer-Verlag, Tokyo, Japan, pp. 211–221.
7. Mang HG, et al. 2004. *Biochim Biophys Acta—Gene Structure and Expression* **1676**: 231–239.
8. Oguchi K, et al. 2004. *Plant Cell Rep* **22**:848–858.
9. Tanaka N, et al. 2005. *J Proteome Res* **4**:1575–1582.
10. Konishi H and Komatsu S. 2003. *Biol Pharm Bull* **26**:401–408.
11. Miche L, et al. 2006. *Mol Plant Microbe Interact* **19**:502–511.
12. Marrs KA. 1996. *Annu Rev Plant Physiol Plant Mol Biol* **47**:127–158.
13. Gonneau J, Mornet R and Laloue M. 1998. *Physiol Plant* **103**:114–124.
14. Kamada M, Higashitani A and Ishioka N. 2005. *Biol Sci Space* **19**:148–154.
15. Bestel-Corre G, et al. 2002. *Electrophoresis* **23**:122–137.
16. Valot B, Gianinazzi S, and Eliane DG. 2004. *Phytochemistry* **65**:1721–1732.
17. Valot B, et al. 2006. *Proteomics* **6**:S145–S155.
18. Valot B, et al. 2005. *Plant Mol Biol* **59**:565–580.
19. Yan S, et al. 2005. *Proteomics* **5**:235–244.
20. Schutzendubel A and Polle A. 2002. *J Exp Bot* **53**:1351–1365.
21. Tripathi BN and Gaur JP. 2004. *Planta* **219**:397–404.
22. Chiang HC, Lo JC, and Yeh KC. 2006. *Environ Sci Technol* **40**:6792–6798.
23. Kawasaki S, et al. 2001. *Plant Cell* **13**:889–905.
24. Roth U, von Roepenack-Lahaye E, and Clemens S. 2006. *J Exp Bot* **57**:4003–4013.
25. Seki M, et al. 2002. *Plant J* **31**:279–292.
26. Salekdeh GH, et al. 2002. *Field Crops Res* **76**:199–219.
27. Reddy ASN. 2004. *Trends Plant Sci* **9**:541–547.
28. Forment J, et al. 2002. *Plant J* **30**:511–519.
29. Requejo R and Tena M. 2005. *Phytochemistry* **66**:1519–1528.
30. Weber M, Trampczynska A, and Clemens S. 2006. *Plant Cell Environ* **29**:950–963.
31. Kung CC, et al. 2006. *Proteomics* **6**:2746–2758.
32. Colditz F, et al. 2004. *Plant Mol Biol* **55**:109–120.
33. Nyamsuren O, et al. 2003. *Physiol Mol Plant Pathol* **63**:17–26.
34. Iturriaga EA, et al. 1994. *Plant Mol Biol* **24**:235–240.

LEAF PROTEOME

Bin Kang, Shuyang Tu, Jiyuan Zhang, and Siqi Liu

16.1 INTRODUCTION

The main functions of leaves are to capture light for photosynthesis, which makes carbohydrates from CO₂ and water. Leaves also perform other functions such as storing food and water and protecting some plants from animals. Generally, a leaf is anatomized as three main structures [1]. As shown in Figure 16.1, the outermost layer of leaf is the epidermis, which is covered by waxy cuticle to prevent water loss. The primary area of photosynthesis in a leaf is in a layer of specialized parenchyma cells called mesophyll. Two types of mesophyll cells are commonly found in a leaf: (a) palisade mesophyll cells with generally rectangular rows located at the upper surface and (b) spongy mesophyll with uneven shapes located at the lower surface. All leaves contain vascular systems that consist of many veins (xylem and phloem) separated from mesophyll by bundle sheath and that function as the transport system for water and minerals throughout the plant. The veins in monocots leaves are parallel to each other, whereas the veins in dicots leaves form branched networks.

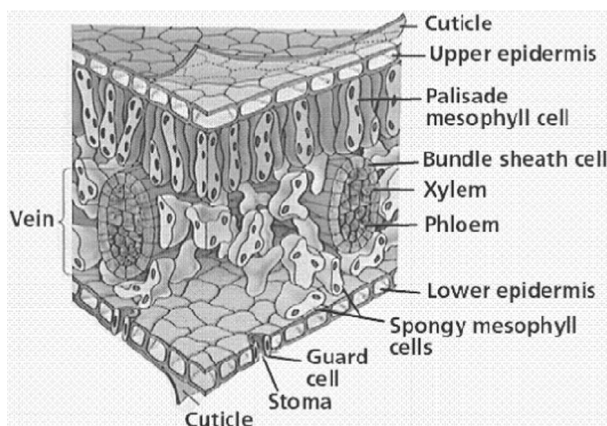


FIGURE 16.1. The anatomic structure of plant leaf.

16.2 PROTEOMICS STUDIES IN PLANT LEAF

Leaf Proteomics: A Hot Area in Plant Proteomics Investigations

Leaves are one of the most important tissues in plants, and they are able to manufacture and store food for plants. The protein components of leaves and their changes have naturally attracted the attention of scientists working in the botanical frontier. Almost right after the emergence of proteomics in the post-genomics era, the protein profiling of plant leaves has been widely investigated and is greatly progressing. Proteomics studies of plant leaves have covered many different plants, from monocots to dicots and from herb to arbor, such as *Arabidopsis*, rice, wheat, maize, citrus, sunflower, clover, cucurbit, beet, holm oak, pea, and tobacco [2–13]. With the completion of genome sequencing (*Arabidopsis*) and the availability of several research tools, the leaf proteome of *Arabidopsis* was systematically surveyed. The leaf proteome of *Arabidopsis* was analyzed with a proteomics strategy of multidimensional shotgun by Weckwerth's group [2]. To avoid interference from high-abundance proteins such as RuBisCO, the protein extracts from *Arabidopsis* leaves were prefractionated with FPLC and were then measured using high-resolution approaches coupled with 2D LC (monolithic silica C18 capillary column) and ESI-MS/MS. A total of 1032 unique proteins were identified from those leaves. 2-DGE was used to compare the protein patterns of the leaves in the *Arabidopsis* ecotypes [14]. Image analysis revealed that the protein patterns of diverse *Arabidopsis* lines were rather similar based upon the 2D gel images. However, two pairs of spots in the ecotype Col-0 were significantly different from those in Ws-2, and they were further identified as the germin-like proteins with the unique different residue, Thr substituted by Lys. On the other hand, despite the lack of genome data for some plant species, many plant leaves have been investigated by proteomics using *de novo* sequencing and BLAST similarity search. For instance, Holm oak is the dominant tree species of most natural communities over large areas of the Western Mediterranean Basin, and it lacks genomic information.

Its leaves were treated with cold TCA/acetone to extract the proteins followed by 2-DGE separation and CBB-G250 staining [11]. The 2D gel images revealed ~400 spots. Out of 43 excised spots, 35 were positively identified as plant proteins, which mainly corresponded to enzymes involved in photosynthesis and energy metabolism. Furthermore, the leaf proteome of Holm oak seedlings grown under greenhouse conditions was partially characterized and compared with that of mature trees in the field [15]. Jorge et al. found that the 2D gel profiles changed according to different tree developmental stages as well as different provenances.

Proteomics Techniques Employed for Studying Plant Leaves

Currently the proteomic technology is still at a prototype phase. There is no unique technique capable of solving all proteomics analysis. A combined strategy is thus often used in this field [16]. In contrast to the proteomics investigations in human and animals, the analytical means in plant proteomics are relatively simple. Although the strategy employing multiple-dimensional chromatographies has been adopted for plant proteomics, the techniques for leaf protein separation still mainly rely on SDS-PAGE and 2-DGE. For the membrane proteins and insoluble fractions, SDS-PAGE is a method of choice. In many cases, 2-DGE is a very popular tool. The IEF strips used in 2-DGE for plant leaf proteomics are of two major types, pH 3–10 and pH 4–7 [5–8], indicating that the majority of leaf proteins possess the apparent pI values—neutral or slight acidic.

It is worth noting that the protocols for extraction of leaf proteins are quite flexible and are dependent on individual labs. Those inconsistent preparations could lead to the different results of proteomic surveys. For example, in the 2-DGE analysis for rice leaf proteome, Kim et al. [17] homogenized the rice leaves in a glass mortar and pestle with a modified LB-TT (see Chapter 11). Komatsu et al. [18] placed the rice leaves in a glass mortar and pestle on ice and extracted the proteins in LB containing 7.5 M urea, 2.5 M thiourea, 12.5% glycerol, 50 mM Tris-HCl, pH 8.0, 2.5% *n*-octylglucoside, 6.25 mM Tris-(carboxyethyl) phosphine hydrochloride, and 1.25 mM protease inhibitor [18], whereas Zhao and colleagues ground the rice leaves to fine powder in a metal pulverizer immersed in liquid nitrogen and allowed the leaf proteins to be precipitated in precooled 10% TCA and 10 mM DTT in acetone for 4 h at -20°C . Finally, the protein pellets were resuspended in the LB containing 7 M urea, 2 M thiourea, 4% NP-40, 10 mM DTT, 0.5% ampholyte (pH 3.5–9.5), 1 mM PMSF, 2 mM EDTA, and 20 mM Tris-HCl, pH 8.5 [3]. With the development of proteomics techniques, the drawbacks of 2-DGE have been gradually revealed (see Chapter 2 for further details). One of the major problems is poor reproducibility caused by numerous and uncertain reasons. If the sample preparations are not conducted by the same protocol, 2-DGE results are expected to be incomparable. Therefore, standardizing the extraction methods, especially for the same plant species becomes necessary.

The Proteomics of Leaf Epidermis. Leaf epidermis covers the entire leaf surface on the top as well as on the underside, and it contains openings called stoma. This tissue functions in prevention of water loss and acts as a barrier to disease organisms from entering the plant. Since leaf epidermis is very thin and presents difficulties

during isolation, its proteomic analysis presents a great obstacle at the preparation step. Karrer et al. [19] developed a method to prepare tomato leaf epidermal cells, and they isolated mRNA from the prepared epidermis, but the amplification of RNA extracted by this method often resulted in failure due to the small amount of RNA obtained. Nakazono et al. [20] developed the laser-capture microdissection (LCM) method to prepare the leaf epidermis. Using an adhesive-coated slide system, large numbers (>10,000) of epidermal cells from ethanol:acetic acid-fixed coleoptiles of maize were obtained by LCM. The RNA extracted from these cells was used to hybridize a microarray chip containing ~8800 maize cDNAs. With limitation of the prepared materials, nevertheless, there is no report on the proteomics analysis of leaf epidermis.

The Proteomics of Leaf Mesophyll. The middle regions of leaves are called mesophyll whose cells are packed with chloroplast, where photosynthesis occurs. Therefore, the proteomics studies have been mainly focused on the chloroplast proteins. Because the proteomics of chloroplast/thylakoid are reviewed in other chapters (see Chapter 23), we do not elaborate on them in this chapter.

The Proteomics of Leaf Vascular System. Plants have a vascular system that carries water and nutrients throughout the plant. These tissues contain a vascular bundle, which may be surrounded by a bundle sheath (layer of parenchyma cells or sclerenchyma cells or both). Technically, isolation of xylem and phloem from the leaf vascular system is not easy, and thus the proteomics investigation to leaf vascular system is currently limited to the dissection of the leaf sheaths. Using detached rice leaf sheaths, proteomics of detached rice leaf sheath responding to GA treatment that usually triggers intermodal or leaf elongation have been performed [18, 21, 22]. 2-DGE was conducted to identify rice proteins whose expression was changed by the GA treatment [21]. Of 352 2D gel spots, 32 proteins exhibited modulation in the expression levels in the GA-treated leaf sheath for 48 h over control [21]. Furthermore, in a preliminary experiment, two techniques, 2D LC and DIGE, were employed to compare the GA-responsive proteomes in rice leaf sheath and resulted in 1248 protein fractions and 1500 proteins identified, respectively [18]. To gain insight into the GA responsive proteins, the cells derived from rice leaf sheath were converted to a constitutive GA response mutant, *slr1* [22]. But further work needs to be carried out before any conclusions can be drawn.

Leaf Senescence and Proteomics

Leaf senescence, visibly observed as leaf yellowing and petal wilting, is the final stage in the development of the leaf and comprises a highly regulated and coordinated activation of self-destruction mechanisms that require expression of specific genes, ultimately resulting in necrosis of the organ. Many researchers regard senescence as a form of PCD on the basis of the following three observations: (i) Cells die at a predictable time and location, (ii) death has some beneficial effect on plant development, and (iii) cell death is encoded in the hereditary material.

Extensive physiological and biochemical studies on leaf senescence in the last three decades have suggested that it is a highly regulated and active process with

involvement of diverse metabolic and biochemical pathway changes [23]. A global view at gene expression would provide invaluable information to explore the senescent mechanisms in plants. Recent comprehensive studies on the transcriptomes of leaf senescence of *Arabidopsis* and wheat revealed that like other developmental processes, leaf senescence was a very dynamic, complex, and active program that required the activation or inactivation of many genes [24, 25]. According to the reports, during the development of leaf senescence, the components of chloroplast, chlorophyll, and other macromolecules underwent a series of degradations, which was a result of several up-regulated genes and whose protein products were responsible for the intensive degradation of proteins, lipids, nucleotides, and polysaccharides.

Studying leaf senescence with proteomics is just at the initial phase. Using TEM, Wilson et al. [26] observed that the most obvious ultrastructural changes during senescence occurred in chloroplasts of white clover leaves with progressive loss of thylakoid integrity and accumulation of osmophilic globules in the stroma. The leaf proteins were separated by 2-DGE and verified by MALDI-TOF-MS. Out of 590 2D gel spots, ~40% of them showed significant senescence-related changes in abundance. Moreover, approximately one-third of those senescence-responsive spots with decreased abundance appeared in the 2D gel profiles of the isolated chloroplasts. TMS analysis demonstrated those spots to be chloroplast proteins, including RuBisCO LSU and SSU, a RuBisCO activase, glutamine synthetase, and the 33-kDa protein of the PS II OEC. Since those proteins are key elements in photosynthesis, the decline in their expression levels emphasized the importance of proteolysis, chloroplast degradation, and remobilization of nitrogen in leaf senescence. Changes in ROS and oxidative stress in leaves are often observed during the plant senescent process. It remains unclear, however, whether the increased ROS are generated actively as the signals to trigger the senescent pathways, or whether ROS are the by-products of the developmental/stress condition that induces senescence. Swidzinski et al. [27] found an increased relative abundance of four catalase isoforms and two forms of mitochondrial MnSOD in senescent *Arabidopsis* cells by 2-DGE. This observation was consistent with other experimental results. For example, transgenic tobacco plants with reduced catalase levels exhibited increased susceptibility to stress conditions and were hyper-responsive to pathogen attack, indicating that the antioxidant enzymes play an important role during the senescence process [27].

Impressively, the proteomics observations have been confirmed by other functional measurements. Using chlorophyll fluorescence spectroscopy, stability of PS I and PS II in rice flag leaves were monitored at two specific peaks, 685 and 735 nm. During leaf senescence, the total fluorescence yields of PS I and PS II decreased significantly, however, the decrease in the chlorophyll fluorescence yield of PS I was greater than in the yield of PS II, suggesting that the rate of degradation in chlorophyll-protein complexes of PS I was greater than in chlorophyll-protein complexes of PS II [28]. On the other hand, photosynthetic CO₂ assimilation rates, carboxylase activity of RuBisCO, chlorophyll and carotenoids content, and the chlorophyll a/b ratio were found at lower levels with progress in senescence. In a proteomics study, thylakoid ATP/ADP carrier (TAAC) was concluded to be localized in the chloroplast (inner) envelope [29]. Its transcription and protein expression were traced during *Arabidopsis*

leaf developmental stages. The amount of the TAAC mRNA was high in developing leaves, decreased during maturation, and increased again during senescence to levels exceeding those present in developing leaves. The change of TAAC at the protein level generally followed its corresponding transcript level. Exposing *Arabidopsis* leaves to various abiotic stress conditions, it was found that the TAAC protein abundance increased during wounding, light stress, oxidative stress, salt stress, and desiccation, whereas it decreased under heat shock conditions. Hence, Thuswaldner et al. [30] hypothesized a role for TAAC in supplying ATP into the thylakoid lumen for mobilization of nitrogen-resources during senescence.

The Modified Proteomes of Plant Leaves

It is well known that plants survive various environmental conditions and possess strong adaptive capacities. PTMs are important for these adaptations [31].

For plant proteins, phosphorylated modification to a protein is one of the most widespread and best understood post-translational event. Using pea leaves as materials, Gomez et al. [32] analyzed the proteins of PS II-enriched thylakoid membrane and detected several phosphorylated proteins, such as PsbA, PsbD, PsbC, PsbH, and two Lhcb. Two phosphorylation sites were confirmed in PsbH, suggesting that PsbH in green plants might have four different phosphorylation states. As compared to the measured rates of phosphorylation/dephosphorylation of PsbH and LHC IIb (Lhcb1), the phosphorylation of PsbH was faster than that of LHC IIb; conversely, the rate of dephosphorylation of LHC IIb was higher than that of PsbH. This led to a model in which the phosphorylation of LHC IIb alone was correlated to the transition between state 1 and state 2. To identify phosphoproteins in rice leaf sheath, Khan et al. [33] incubated the protein extracts with [γ - 32 P]ATP, visualized the phosphorylated proteins with silver staining of 2D gel, and exposed 2D gels on X-ray film. In this preliminary experiment, ~5% of the 2D gel spots were found as phosphorylated rice proteins; however, these data need to be reevaluated. When rice seedlings were treated with various hormones and stresses, MDH was consistently phosphorylated in rice leaf sheath, leaf blade, and root, indicating that the phosphorylation cascade seemed to be important for glycolytic metabolism processes and calcium-signaling in response to environmental stimuli [33].

As in other eukaryotes, the carbohydrate moieties of glycoproteins in plants participate in many biological processes. N-linked glycans are the most conserved forms of protein glycosylation in plants; by contrast, O-linked glycans tend to be a wide and disparate group of modifications in plants. In spite of a number of reports on individual glycoproteins in plants, the information on plant glycoproteomes are largely unknown. Balen et al. [34] first attempted a glycoproteomic survey in *M. gracillis* tissues. The proteins extracted from different tissues, such as shoot, callus, and hyperhydric regenerant, were separated by 2-DGE and transferred to nitrocellulose membranes followed by deglycosylation with PNGase F. The 2D gel spots corresponding with deglycosylation and the released oligosaccharides were analyzed by MALDI-TOF/TOF-MS. The electrophoretic and PNGase F-digested results indicated that the *N*-glycosylation patterns were clearly dependent on plant tissue. For instance, the glycosylated spots

in callus were significantly more than those of shoot even though a 42-kDa glyco-protein was shared by the two tissues. MS analysis revealed that the oligosaccharides in shoot and callus were similar, releasing the common glycan mass signals at Hex2dHex1HexNAc2 structure, as $[M + Na]^+$ and $[M + 2Na-H]^+$ at m/z 917.12 and 939.10, and at Hex3dHex1HexNAc2 structure, as $[M + Na]^+$, $[M + 2Na-H]^+$, and $[M + Na + K-H]^+$ at m/z 1078.96, 1100.94, and 1116.92, respectively.

NO is a crucial regulator in plant physiological processes such as stomatal closure, growth, and development, and it is also an important messenger in plant defense signaling against microbial pathogens. One of the NO effects to cellular components is the modification of the amino residues of proteins, in which Cys residue is a primary target in this modification via oxygen-dependent chemical reactions or the transfer of NO from a nitrosothiol to a protein sulfhydryl group. *S*-Nitrosylation is believed to be an important modification event that is involved in a number of cross-talking signal transduction pathways in plants. *Arabidopsis* plants were cultivated in a growth chamber and treated with NO concentrations of 1250 mL/L for 10 min under light followed by harvest of the leaves for proteomics analysis using a combination of biotin affinity and nLC-MS/MS [35]. A total of 52 proteins from the leaves were identified, which represented candidates for *S*-nitrosylated products, including stress-related, redox-related, signaling/regulating, cytoskeleton, and metabolic proteins. Of those proteins, 19 were previously reported as targets of *S*-nitrosylation in animals. Further experiments demonstrated that the *S*-nitrosylated modification could lead to inactivate the enzymatic activity. Incubation of 1 mM of GSNO with the leaf GAPDH reduced 90% of its activity, whereas 1 mM SNP inhibited ~50% enzyme activity of GAPDH. Moreover, the inhibition could be rescued with addition of high concentration of DTT, indicating that the modification on these sulfhydryl groups was reversible.

Rice Leaf Proteomes

Rice is one of the most important cereals on our planet, and it provides the food supply to 60% population worldwide. Since the process of photosynthesis takes place in the rice leaf, how solar radiation is effectively transformed into biomass through rice leaves is a key issue in rice research. Recently, proteomics investigation on rice leaves has been carried out by a number of laboratories (Figure 16.2 shows an example of 2-DGE separation of rice leaf proteins) with some success.

An excellent study is the monitoring of the protein expression profiles of rice leaves during its growth phases using 2-DGE, by Zhao et al. [3]. Differential protein expressions among the six phases were analyzed by image analysis, which allowed the identification of 49 growth-dependent 2D gel spots. Those spots were further verified by MALDI-TOF/TOF-MS, in which 89.8% of them were confirmed to be rice proteins. Some of the interesting rice proteins were further confirmed with immunoblotting. On the basis of these observations, the authors concluded the following: (i) Protein expression in rice leaves, at least for high or middle abundance proteins, was attenuated during growth, especially some chloroplast proteins; however, the change was slow and the expression profiles were relatively stable during

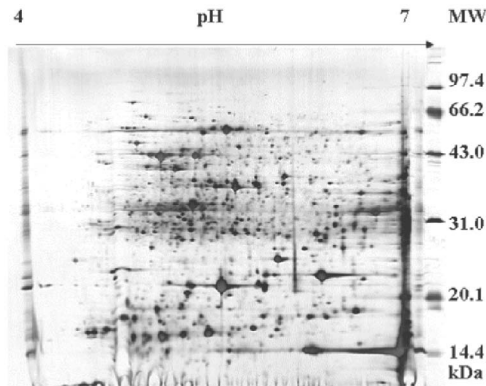


FIGURE 16.2. Separation of rice leaf proteins by 2-DGE and silver nitrate staining. The work is accomplished by proteomics group of Beijing Genomics Institute, Chinese Academy of Science.

rice development. (ii) RuBisCO was expressed at constant levels at different growth stages accompanied with a high ratio of the degraded LSU. (iii) The expression of antioxidant proteins, such as SOD and POX, declined at the early ripening stage.

The responses of rice leaf proteins to the temperature stress were studied by Cui et al. [36]. The seedlings were exposed to three temperatures at 5°C, 10°C, and 15°C, and the leaves were collected for proteomics survey. Approximately 1700 2D gel spots were well-resolved with 60 up-regulated protein spots responding to the progressively low temperature stress. These cold-responsive proteins of rice leaves included four involved in protein biosynthesis, four molecular chaperones, two proteases, eight involved in the biosynthesis of cell wall components, and seven involved in antioxidative/detoxifying events and signal transduction. Using TargetP program, the subcellular localization of these identified proteins revealed 43% of them as chloroplast-localized. It was suggested that the leaf proteins such as chaperones, proteases, and cell wall components could play the important roles in tolerance to cold stress, as well as that the chloroplast proteins were sensitive to low temperature as well.

Under salt stress, the rice leaf proteomes exhibited changed protein profiles. Based upon the Kim et al. [17] report, seedlings were hydroponically cultured for 1 week in distilled water followed by replacement of distilled water with the nutrient solution, containing 40 mg each of N, K, C, and Mg, 10 mg of P, 0.5 mg of Mn, 0.05 mg of Mo, 0.2 mg of B, and 0.01 mg each of Zn and Cu, respectively. After 16 days, 10 uniformly grown seedlings were carefully selected for treatment with salt. A comparison of 2D gel protein profiles of the third leaves between the untreated control and salt-stressed leaves revealed 55 differential expressed spots, including 47 up-regulated spots. Of those salt-responsive spots, the identity of 33 protein spots was determined by nESI-LC-MS/MS. Most of the identified proteins belonged to the major metabolic processes such as photosynthetic CO₂ assimilation and photorespiration, suggesting a good correlation between salt-stress-responsive proteins and leaf

morphology. Moreover, 2D immunoblot analysis and enzymatic activity determinations for the third leaves demonstrated remarkable changes in the enzymes associated with oxidative damage.

16.3 CONCLUSIONS

The proteomics investigation in plant leaves has indeed provided rich information to further explore the physiological mysteries at molecular level. The overview of protein expression in many plant leaves demonstrated that the proteins involved in photorespiration and metabolic process were obviously regulated in their expression as responding to various stimuli. The leaves of two plants, *Arabidopsis* and rice, have been extensively studied at the level of the proteome, becoming good models for leaf proteomics in other species. We have to realize, nevertheless, that the currently proteomics achievements are still primary because the information is too scattered to be integrated into an entire feature for elucidation of physiological functions of leaves. Development of proteomics technology and establishment of cell models for plant leaves are urgently required to discover the protein changes, including abundance, activity, configuration, modification, and translocation, which are tightly associated with the behaviors of the plant physiology.

16.4 FIVE-YEAR VIEWPOINT

In next five years, we have full confidence for the proteomics technology, growing from prototype to a mature status. The proteomics of the future will not only cover the areas for protein separation and identification, but will also concern other emerging techniques, such as protein chips and protein imaging. Undoubtedly, these advances will take plant proteomics to new heights in the coming years. Specifically for leaf proteomics, in our opinion, three research fields will have to be tackled for bridging the gap between proteins and leaf physiology. First of all, chloroplast as an elementary organelle in the leaf, along with its protein components, has great biological significance. Although the chloroplast proteomes have been studied by many laboratories, the dynamics and precise analysis has not yet been accomplished. In particular, protein expression profiling in the chloroplast under different growth phases or stimuli remains to be done. Second, we will need to engineer stable and diverse leaf cell cultures. Due to the complex composition of cells in leaf, a theoretically cellular model will be useful to elucidate the molecular mechanisms upon the proteomics data. Finally, as reported, the leaf proteomes have not exhibited the dramatic alterations during plant growth. This observation raises a question whether the modification forms of the leaf proteins play pivotal roles related to the leaf functions.

ACKNOWLEDGMENTS

This work was supported by Grant (KSCX2-YW-N-026) from the Chinese Academy of Sciences and National Natural Science Foundation of China (30221004).

REFERENCES

1. Stern K, Bidlack J, and Jansky S. 2007. *Introductory Plant Biology*, 11th edition, McGraw-Hill, New York.
2. Wienkoop S, et al. 2004. *Rapid Commun Mass Spectrom* **18**:643–650.
3. Zhao C, et al. 2005. *Proteomics* **5**:961–972.
4. Bahrman N, et al. 2004. *Proteomics* **4**:2672–2684.
5. Riccardi F, et al. 2004. *Plant Physiol Biochem* **42**:1003–1011.
6. Gancel AL, et al. 2006. *J Agric Food Chem* **54**:6212–6218.
7. Garcia JS, et al. 2006. *J Agric Food Chem* **54**:8623–8630.
8. Wilson KA, et al. 2002. *Proteomics* **2**:1114–1122.
9. Kamada M, Takaoki M, and Ishioka N. 2004. *Biol Sci Space* **18**:156–167.
10. Hajheidari M, et al. 2005. *Proteomics* **5**:950–960.
11. Jorge I, et al. 2005. *Proteomics* **5**:222–234.
12. Schiltz S, et al. 2004. *Plant Physiol* **135**:2241–2260.
13. Dani V, et al. 2005. *Proteomics* **5**:737–745.
14. Schlesier B, et al. 2004. *Phytochemistry* **65**:1565–1574.
15. Jorge I, et al. 2006. *Proteomics* **6**:S207–S214.
16. Chen S and Harmon AC. 2006. *Proteomics* **6**:5504–5516.
17. Kim DW, et al. 2005. *Electrophoresis* **26**:4521–4539.
18. Komatsu S, Zang X, and Tanaka N. 2006. *J Proteome Res* **5**:270–276.
19. Karrer EE, et al. 1995. *Proc Natl Acad Sci USA* **92**:3814–3818.
20. Nakazono M, et al. 2003. *Plant Cell* **15**:583–596.
21. Shen S, Sharma A, and Komatsu S. 2003. *Biol Pharm Bull* **26**:129–136.
22. Tanaka N, et al. 2005. *J Proteome Res* **4**:1575–1582.
23. Buchanan-Wollaston V, et al. 2003. *Plant Biotechnol J* **1**:3–22.
24. Lin JF and Wu SH. 2004. *Plant J* **39**:612–628.
25. Gregersen PL and Holm PB. 2007. *Plant Biotechnol J* **5**:192–206.
26. Wilson KA, et al. 2002. *Proteomics* **2**:1114–1122.
27. Swidzinski JA, Leaver CJ, and Sweetlove LJ. 2004. *Phytochemistry* **65**:1829–1838.
28. Tang Y, Wen X, and Lu C. 2005. *Plant Physiol Biochem* **43**:193–201.
29. Ferro M, et al. 2003. *Mol Cell Proteomics* **2**:325–345.
30. Thuswaldner S, et al. 2007. *J Biol Chem* **282**:8848–8859.
31. Rossignol M, et al. 2006. *Proteomics* **6**:5529–5548.
32. Gomez SM, et al. 2002. *Mol Cell Proteomics* **1**:46–59.
33. Khan M, Takasaki H, and Komatsu S. 2005. *J Proteome Res* **4**:1592–1599.
34. Balen B, et al. 2006. *J Proteome Res* **5**:1658–1666.
35. Lindermayr C, Saalbach G, and Durner J. 2005. *Plant Physiol* **137**:921–930.
36. Cui S, et al. 2005. *Proteomics* **5**:3162–3172.

ANTHER PROTEOME

Nijat Imin

17.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

In higher plants the development of the male gametophyte is a well-programmed and elaborate process that plays a crucial role in plant reproduction. Male gametophyte development takes place inside a highly differentiated male gametophytic organ, the stamen. A stamen is the male reproductive organ of plants and consists of an anther borne on a slender filament. The filament is a tube of vascular tissue that connects the anther to the flower and serves as a conduit for water and nutrients. Each anther has four elongated macrosporangia, or anther loculi. Four anther wall layers (the epidermis, the endothecium, the middle layer, and the tapetum) enclose a locule where the microsporegenesis took place (Figure 17.1). During anther development in angiosperms, a series of processes generates a set of cells that serve microspore production and dispersal. In the early stage of development, the anther consists of a mass of undifferentiated meristematic cells, which are surrounded by a partially differentiated epidermal cell layer. As development progresses, the archesporial cells appear at the four corners of the young anther and further differentiate into primary parietal cells and primary sporogenous cells. Later the primary parietal cells differentiate into

A . Photo of a rice anther at the middle microspore stage. The arrow indicates the plane of the cross section shown in B.

B. Microscopic view of cross section of an anther at the middle microspore stage.

C . Anther developmental stages defined by microspore development.

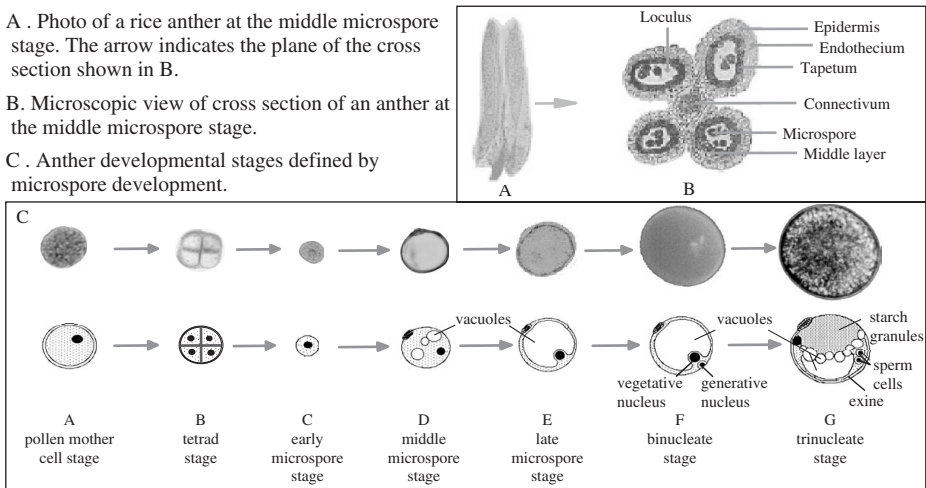


FIGURE 17.1. Schematic diagram showing anther and its developmental stages.

endothecium, middle layer, and tapetum that envelop the core of the anther locule, whereas the primary sporogenous cells develop into diploid pollen mother cells (PMC). The innermost tapetum layer of the anther locule provides the sporogenous cells with enzymes, nutrients, and structural materials necessary for development. After meiosis, tetrads of haploid microspores are formed that subsequently become separated as individual pollen grains. The developing young microspore undergoes an asymmetric mitotic division, forming a large vegetative cell and a small generative cell that is enclosed within the vegetative cell. This bicellular structure is called the pollen grain. The generative cell inside the pollen grain forms two identical sperm cells by a round of symmetrical mitotic division. At the same time, the stomium differentiates and, eventually, together with the endothecium and the septum, degenerates (anther dehiscence) to release the pollen. The whole developmental process of microgametogenesis is controlled by coordinated gene expression in both somatic and gametophytic cells.

The last two decades have been marked by increasing efforts to decipher the genetic and molecular basis of anther development and functions [reviewed in references 1–5]. Temporal and spatial regulation of gene expression has been reported during male gametophyte development in higher plants [6–9] and its response to abiotic stresses [10]. Microarray analysis has also been utilized to dissect anther development [10–12]. In the 1990s, some attempts were made to separate anther proteins using 2-DGE in *lilium* [13] and *rice* [14]. However, little protein identification has been done from these 2-DGE procedures. Professor Barry Rolfe's group at the Australian National University (Australia) has been working on the anther proteomes since 2000, combining high-resolution 2-DGE and MS analysis. This combination has proven to be a powerful tool in identifying new proteins involved in anther development and its response to environmental stresses [15–19]. Here I describe anther proteome and its dynamic change during anther development and response to abiotic stresses in

plants and compare proteomic approaches with transcriptomics approaches in terms of correlation of accumulation, advantages, and disadvantages of both methods.

17.2 METHODOLOGY AND STRATEGY

A key prerequisite for studying the anther proteome is to obtain synchronous sample populations. For instance, in rice, one of the main difficulties of working with anthers is the lack of synchrony among developing anthers. Thus, a method needed to be devised which provides anthers belonging to certain developmental stage(s) can be predicted. Parameters such as anther length, days before dehiscence, and cytology of anthers and microspores can be used to identify exact stages of anther development. In rice, auricle distance, the distance between the auricle of the flag leaf and that of the penultimate leaf, is commonly used as a nondestructive measurement to gauge the stage of development of the rice anthers. In rice, anthers can be removed from immature panicles with precision forceps and scalpels under a dissection microscope. Special care must be taken not to include pistil parts such as stigma and style or any other tissues. Although anthers are rich in proteins (compared to other plant tissues), depending on species, one might need a large number of anthers to get enough protein sample for proteomic analysis. For example, approximately 1000 anthers at the early stages are required for analytic gels in rice. For protein identification, one might need to load five times more sample onto preparative gels. Anther samples should also be collected by snap freezing in liquid nitrogen to avoid protein degradation. The most common method used for anther protein extraction is using 10% TCA in acetone and subsequent washings in acetone. This method not only denatures proteins but also removes other interfering compounds such as DNA, RNA, pigments, and phenolic compounds. For solubilization, 4% CHAPS in 9 M urea LB works well. High-resolution IPG strips (pH 3–10, pH 4–7, and pH 6–11) are also a necessity for separating thousands of proteins. Narrow-pH-range IPG strips (one pH unit) have also proven to be very effective in resolving more proteins (generally about 10% increase compared to wider IPG strips). The longer the strip, the larger the amount of protein that can be loaded and the better the resolution for total proteins. Typically, 18 to 24-cm-long IPG strips are used to display anther proteomes. For SDS-PAGE runs, large format gels are recommended. For protein detection, silver stain seems to work very well, because it is much more sensitive than colloidal CBB staining. SYPRO Ruby staining is also very sensitive. However, not one type of stain will detect all proteins (Figure 17.2). Recently, 2D-DIGE system has proven to be a better option due to its sensitivity and advantages in data analysis. Image analysis with Melanie software (GeneBio, Geneva, Switzerland) has been used successfully.

In the early years, Edman sequencing of proteins that are transferred to PVDF membranes was an option for protein identification. However, due to the introduction of high-throughput MS analysis, it became much less common and MS analysis has taken its place as the main tool of protein identification. While MALDI-TOF-MS is the first choice for protein identification, MS/MS analysis yields much more valuable

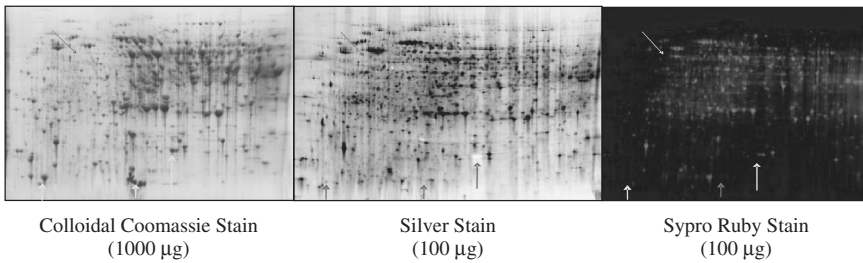


FIGURE 17.2. Comparison between colloidal CBB staining, silver staining, and SYPRO Ruby staining. Red arrows indicate protein spots that are not detected or negatively stained, and yellow arrows indicate properly stained protein spots. See insert for color representation of this figure.

data and increases the level of confident matches. Anther protein identification is typically carried out by N-terminal Edman sequencing on the Procise 494-01 sequencer system (Perkin–Elmer Life Sciences, Boston, MA) and by MS on Micromass ToFSpec 2E–TOF–MS (Manchester, UK) or PE SCIEX (Foster City, CA), QSTAR hybrid LC–MS/MS Q–TOF, or Micromass LC–MS/MS Q–TOF systems. Search engines such as Profound, MassLynx, and Mascot work well for both MS data and MS/MS data. However, a key component of the searches is the databases. Besides general nonredundant database, species-specific EST databases or any other specific databases are needed for high success rates in protein identification. A flow chart for the analysis of anther proteomes by 2-DGE/MS approach is shown in Figure 17.3.

Proteome Analysis of Anther Development

The anther is by far the best-characterized by proteomics [20], for which high-quality 2D gel reference maps have been generated during male gametophyte development in the rice cultivar Doongara under strictly controlled environmental conditions [15, 16]. Anthers at the young microspore stage (from tetrad stage to early microspore stage) were separated by 2-DGE, and silver staining detected over 4000 spots in the pH range of 4–11 that represent approximately 10% of the estimated total genomic output of rice. Qualitative and quantitative analysis of differentially displayed spots was performed with Melanie 3 software. Two-hundred and seventy-three spots, collected from either PVDF membranes or colloidal CBB-stained 2D gels, were analyzed by Edman sequencing and MALDI–TOF–MS or Q–TOF–LC–MS/MS. MS analysis was performed with the search engine Profound and MassLynx to search nonredundant database and rice EST database. A total of 53 spots were identified that represented 43 different proteins. Of these 43 proteins, 37 proteins matched to rice ESTs, and 22 proteins were functionally assigned. The major protein groups identified were housekeeping proteins, glycine-rich proteins, translationally controlled tumor protein (TCTP, a first report from rice anthers), chaperones (HSP60 and HSP70), proteins with protective functions GST, and APX, members of the 20S proteasome, and the Jun activation domain-binding protein. One of the interesting characteristics

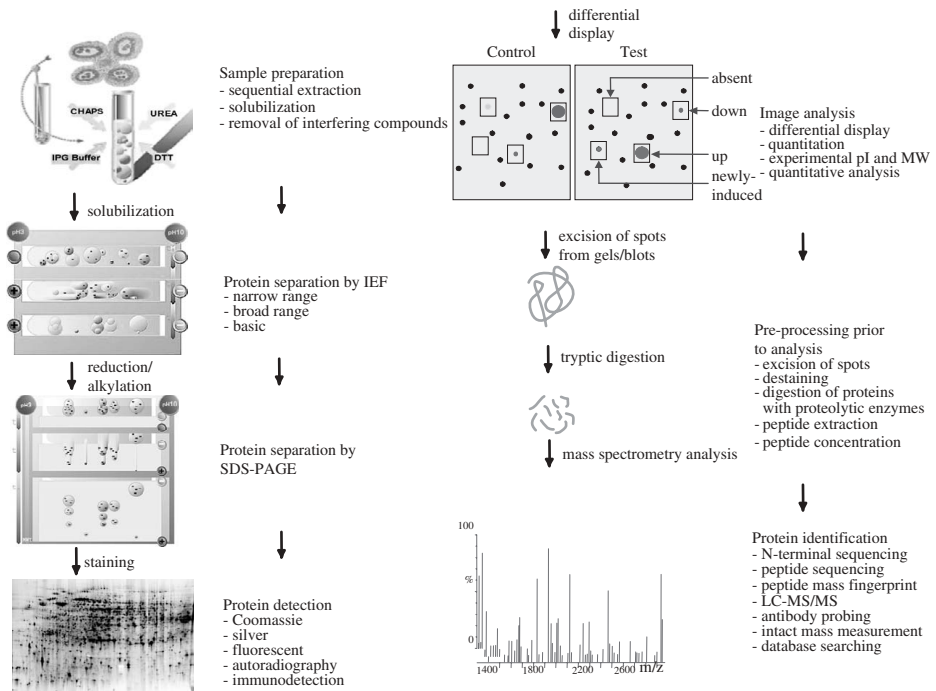


FIGURE 17.3. Typical flow chart for the analysis of anther proteomes by 2-DGE/MS approach. Anther proteins are extracted, solubilized, and separated by 2-DGE, after which they are visualized and differentially displayed and selected for identification. The protein spots are excised and digested by a site-specific protease such as trypsin. Peptides subsequently are extracted, and a small fraction is used to determine the peptide ion masses by MALDI-TOF-MS. The experimental peptide ion masses are then searched against proteins predicted from genomic sequence data. If no positive identification can be achieved, (partial) sequences of the remaining peptides are determined by MS/MS. The sequence tags (partial sequence, together with the parent and fragment ion masses) are used to search against EST or annotated genomic sequence data for protein identification.

of anther proteome is that the existence of proteins with cell protection functions (about 20% of the identified proteins) such as HSPs, PDI, 20S proteasome, enolase, and APX suggests an important role in the microspore development and indicates the presence of a preformed defense and protection against such factors even in the absence of environmental stresses. In the other study, a detailed analysis of the six discrete pollen developmental stages has been provided with an aim to investigate the changing patterns of protein synthesis in anthers [16, 17]. The stages investigated are PMC stage and tetrad stage, early young microspore stage, uninucleate microspore stage, binucleate microspore stage, and trinucleate pollen grain stage. Extracted proteins were separated by 2-DGE (pH 4–7) and were stained with silver nitrate to resolve ~2,500 spots by image analysis (Melanie 3) to produce

proteome maps at the six different development stages. These proteome maps were compared, where ~155 spots were found to change more than twofold. The 155 differential spots were excised from colloidal CBB stained gels, and the proteins were analyzed by MALDI-TOF-MS, resulting in the identification of 40 (26%) proteins by PMF. These proteins represented 33 unique gene products that were mainly involved in carbohydrate and lipid metabolism, signal transduction, cytoskeleton and cell-wall formation, nucleotide and amino acid metabolism, and stress responses (Table 17.1). Moreover, four proteins that could not be identified by PMF analysis were analyzed by Edman sequencing to identify multiple-charge isoforms of vacuolar acid invertases, fructokinase, β -expansin, and profilin. With the establishment of rice anther proteome and identification of large number of proteins, the global protein expression profiles of male gametophyte development in rice is now better understood.

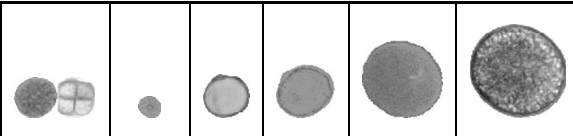
Establishment of Anther Proteome Reference Maps

Comprehensive proteome reference maps of rice anthers have been constructed according to the SWISS-2DPAGE standards and are available for public access at <http://semele.anu.edu.au/2d/2d.html>. The dedicated website contains clickable 2D gel images over the pH range of 4–7 (general) and 6–11 (basic). It is a federated 2-DGE database and is integrated into the Swiss-Prot database.

Anther Proteome and its Response to Biotic and Abiotic Stresses

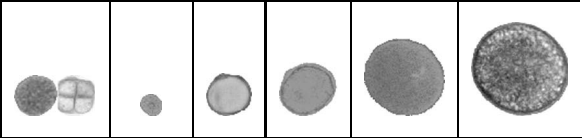
Low temperature has a detrimental effect on male reproductive development in rice, and 2-DGE profiling of the rice anther proteome has been able to identify cold-responsive proteins that may be involved in cold-induced male sterility [19]. A semi-dwarf indica-type cold-sensitive cv. Doongara along with a relatively cold-tolerant Hungarian cv. HSC55 was grown in a controlled growth chamber (30/20°C day/night cycle followed by 12/12 h day/night cycle at the time of panicle initiation to maximize cold stress effect) until anthers were ready to be collected. After 2 and 4 days of cold treatment, 33% and 86% sterility, respectively, were observed in Doongara anthers, compared to just 24% sterility in HSC55. The cold-treated anthers also showed considerable cytological abnormalities in Doongara. 2-DGE was carried out on 18 cm IPG strips (pH 4–7 and 6–11) for the first dimension followed by the second dimension separation on 12–14% Excel Gels (GE Healthcare Bio-Sciences AB). Visualization of spots on analytical and preparative 2D gels was done with silver nitrate and colloidal CBB G-250, respectively, followed by image analysis with Melanie 3 software. The differentially displayed protein spots were classified on the basis of their twofold change in percentage volume (% vol) between untreated and cold-treated plants in at least three replications. Approximately 4000 silver nitrate-stained protein spots were reproducibly resolved over a combined pH range of 4–11. Results revealed changes in abundance of 37 anther protein spots after 1, 2, and 4 days of cold-treatment in Doongara. Thirty-two proteins were up-regulated, whereas four proteins were down-regulated. Two anther-specific proteins (putative LTP and Osg6B) and

TABLE 17.1. Some of the Proteins that Are Differentially Displayed During the Anther Development

Protein Identities						
	Spot Regulation					
	PMC	EYM	MM	LM	BN	TN
Vacuolar acid invertase	—	—	—	+	+	+++
Putative fructokinase II	+	+	+	++	+++	+
Submergence-induced protein 2A	++	++	++	++	++	+
Jab1 protein	+	+	+	+	+	++
Putative reductase	—	—	—	—	—	+
Putative beta-galactosidase	—	—	—	+	++	+
Glutathione <i>S</i> -transferase II	+	+	+	+	++	++
Putative-soluble inorganic pyrophosphatase	—	—	—	—	+	+
Beta 1 subunit of 20S proteasome	++	++	+	++	++	+
Nucleoside diphosphate kinase	+	+	+	++	++	++
Mitochondrial processing peptidase alpha-II chain	+	+	+	+	++	++
Protein disulfide isomerase	+	++	++	++	+	+
Translationally controlled tumor protein	+	+	+	++	++	+
Beta-ketoacyl-ACP synthase	+	+	++	+	+	+
GSH-dependent dehydroascorbate reductase 1	+	+	+	+	++	+
Putative glutathione <i>S</i> -transferase	++	++	++	++	+	+
Profilin	—	—	+	+	++	+++
Resistance complex protein I2C-2-like protein	+	+	+	++	++	++
Beta expansion	—	—	—	+	+	+++

(continued)

TABLE 17.1. (Continued)

Protein Identities						
	Spot Regulation					
	PMC	EYM	MM	LM	BN	TN
Putative cholinephosphate cytidylyltransferase	—	—	+	+	+	—
Aminotransferase-like protein	+	+	+	+	++	++
Putative legumin-like protein	—	—	—	+	++	+
Nonspecific lipid transfer protein	+++	+++	++	+	—	—
Lipase-like protein	+	+	—	—	—	—

Source: Courtesy of Kerim et al. [17].

“—” indicates that the protein spot is absent, “+” indicates that the spot is present, “++” indicates more than twofold increase, “+++” indicates more than fivefold increase. PMC, pollen mother cell stage (combined with tetrad stage); EYM, early young microspore stage; MM, middle microspore stage; LM, late microspore stage; BN, binucleate stage; TN, trinucleate stage.

a calreticulin were down-regulated and a cysteine synthase, a β -6 subunit of the 20S proteasome, an H protein of the glycine cleavage system, cytochrome c oxidase subunit VB, an osmotin protein homologue, a putative 6-phosphogluconolactonase, a putative adenylate kinase, a putative cysteine proteinase inhibitor, a ribosomal protein S12E, a CCOMT, and a mono-DHAR were up-regulated. Accumulation of these proteins did not vary greatly after cold treatment in panicles of cv. Doongara or in the anthers of the cv. HSC55. These proteins have potential roles in protein synthesis and folding, stress responses, lipid biogenesis, cell-wall formation, protein breakdown, and energy metabolism. A novel protein with unknown function, named OsCIA (cold-induced anther protein), was identified in this study. The expression of the OsCIA protein was also detected in panicles, leaves, and seedling tissues under normal growth conditions. Interestingly, though the OsCIA mRNA was present in anthers, panicles, leaves, and seedlings of Doongara, its expression did not change with cold treatment, indicating OsCIA protein accumulation is regulated at the post-transcriptional level. Two potential roles for OsCIA in anther development were proposed. OsCIA may function as a part of a negative regulatory pathway (repressor of anther development) for normal anther development and also as a chaperone to protect cells during normal plant development as well as under environmental stresses.

To understand the consequence of low-temperature effect on male reproductive development in rice, anther proteins at the trinucleate stage, with or without cold treatment for 4 days at 12°C at the young microspore stage, were extracted, separated by 2-DGE (24-cm IPG strips with pH range of 4–7 used), compared in the cold-sensitive cultivar Doongara, and analyzed with Melanie 3 software. Out of a total of 3000 spots on the silver-stained 2D gels, 70 were differentially expressed after cold stress. Among these, 12 spots were newly induced, 47 were up-regulated, and 11 were down-regulated. Sixty-five differentially expressed spots were excised from colloidal CBB-stained gels and digested with trypsin, followed by analysis by MALDI–TOF–MS. Eighteen proteins (26%) were identified by PMF analysis, and the majority were matched to APX that included up- and down-regulated protein spots. APX is an important antioxidant enzyme that is involved in the detoxification of ROS by scavenging the H₂O₂ that was generated in plants during metabolism and under stress. Its involvement in the tolerance of the cell to low-temperature stress in rice has also been reported [21]. Other identified proteins included glycogen phosphorylase, HSP70, aldolase C-1, UDP-glucose pyrophosphorylase, a putative 2, 3-bisphosphoglycerate-independent phosphoglycerate mutase, beta expansin, and a nucleotide diphosphate kinase (NDPK) 1. These seven proteins were also found to have much smaller molecular masses when compared with those of the matched proteins, suggesting these proteins to be partially cleaved or degraded. Thus, 2-DGE profiling of the rice anther proteome has been able to demonstrate that low-temperature stress at the young-microspore stage enhances and induces selective partial degradation of proteins in rice anthers at the trinucleate stage. However, further studies are needed to examine and to understand this phenomenon of partial protein degradation after cold treatment.

Proteomics has also been applied to investigate cytoplasmic male sterility in rice [22]. The proteins of HL-type cytoplasmic male sterility (CMS) rice anther of YTA (CMS) and YTB (maintenance line) were separated by 2-DGE with pH 3–10 nonlinear gradients as the first dimension and SDS-PAGE as the second dimension. The silver-stained proteins spots were analyzed using Image Master software. There were about 1800 detectable spots on each 2D gel, and about 85 spots were differential-expressed. With direct MALDI–TOF–MS analysis and protein database searching, nine protein spots out of 16 were identified. Among those proteins, there were putative nucleic acid binding protein, ADP-glucose pyrophosphorylase (AGPase), UDP-glucuronic acid decarboxylase, putative calcium-binding protein annexin, putative acetyl-CoA synthetase, and putative LADH. These are closely associated with metabolism, protein biosynthesis, transcription, signal transduction, and so on, all of which are cell activities that are essential to pollen development. AGPase catalyzes a very important step in the biosynthesis of alpha 1,4-glucans (glycogen or starch) in bacteria and plants. The lack of the AGPase in the male sterile line might directly result in the reduction of starch, and the synthesis of starch was the most important processes during the development of pollen. The reduction of putative LADH and putative acetyl-CoA synthetase is seemingly involved in pollen sterility in rice.

This study gave new insights into the mechanism of CMS in rice and demonstrated the power of the proteomic approach in plant biology studies.

Comparison of Proteomics and Transcriptomics Profiling of Rice Anthers

Professor Koike's group at the National Agricultural Research Center, Japan has also been working on cold-induced male sterility in rice, and they have recently reported a microarray analysis of rice anther genes under chilling stress at the microsporogenesis stage [10]. They treated cold-sensitive cv. Hayayuki at 12°C for 5 days for the microarray analysis. The gene expression profile during the microspore development process under chilling stress was revealed using a microarray that included 8987 rice cDNAs. As many as 160 cDNAs were up or down-regulated (at least twofold) by chilling during the microspore release stage. This microarray analysis should be comparable with proteomic analysis of cv. Doongara, since both rice cultivars are cold-sensitive and conditions used are very similar. However, only expression of a few transcripts (cysteine synthases and APX) seems to correlate positively with the accumulation of actual protein products. It is not surprising, given the fact that there is poor correlation between RNA levels and protein accumulation [23, 24]. A symbolic example of this poor correlation comes from *OsCIA*, the transcript of which does not change upon low-temperature treatment, in rice anthers. That is why any transcript analysis tools will not be able to pick up this gene as differentially displayed. However, only the protein accumulation is altered by cold treatment, indicating post-transcriptional regulation of this gene. This demonstrates an advantage of proteomics analysis over transcriptomics analysis.

17.3 CONCLUSIONS

Notable progress has been made in the field of anther proteome research in the past few years. This is mostly due to developments in sample preparation, high-resolution protein separation, advance in MS for protein characterization, and advanced bioinformatics. The aim of current anther proteomics is not limited to simple cataloguing of protein species, but extended to qualitative and quantitative analysis of proteins in functional biological context. Integration of plant proteomics and systems biology will enhance the understanding of complex signaling and metabolic networks underlying anther development and interaction with the environment and will improve crop yield and plant tolerance to environmental stresses.

17.4 FIVE-YEAR VIEWPOINT

The major issue impacting 2-DGE, in the context of proteomics, is the fact that it is not possible to array/identify all proteins in a sample on a single gel. Despite the advances made in 2-DGE technology in recent years, including improvements in protein solubilization and application of narrow range IPG strips, there are still a number

of protein groups such as low-abundance proteins, hydrophobic proteins, basic proteins, small proteins and peptides, or large proteins that are not well-represented on 2D gels. PTMs are also underrepresented in current studies. Nevertheless, emerging technologies in proteome analysis such as chip-based methods, “molecular scanner,” ICAT peptide labeling and MudPIT, and advancement in MS will improve the sensitivity of proteomics and will contribute to achievement of a “near-to-complete” proteome in the near future. LCM allows you to capture single cells or sections of cells and then analyze them using the normal nucleic acid and protein analysis methods. LCM has been applied in plant systems to analyze transcriptomes [25, 26]. LCM and MS enable large-scale protein profiling of captured tissue samples and have already been used in animal research [27, 28]. It is essential to capture different layers of anther walls and developing microspores at critical stages and profile their proteome compositions, which will ultimately unravel the complexity and dynamics of anther development and its role in pollen maturation and dispersal. Anther is very susceptible to abiotic stresses which lead to pollen sterility. Effects of low temperature on anther proteome have been profiled extensively. Investigation into the impact of other forms of stresses such as heat, drought, radiation, and nutrient deficiency on the anther proteome will also point to the molecular causes of anther susceptibility and stress-induced male sterility.

REFERENCES

1. McCormick S, et al. 1991. *Symp Soc Exp Biol* **45**:229–244.
2. Goldberg RB, Beals TP, and Sanders PM. 1993. *Plant Cell* **5**:1217–1229.
3. McCormick S. 2004. *Plant Cell* **16**:S142–153.
4. Sterling C. 1963. *Biol Rev Camb Philos Soc* **38**:167–203.
5. Vasil IK. 1967. *Biol Rev Camb Philos Soc* **42**:327–373.
6. Koltunow AM. 1990. *Plant Cell* **2**:1201–1224.
7. Scott R, et al. 1991. *Plant Mol Biol* **17**:195–207.
8. Tsuchiya T, et al. 1992. *Plant Mol Biol* **20**:1189–1193.
9. Tsuchiya T, et al. 1994. *Plant Mol Biol* **26**:1737–1746.
10. Yamaguchi T, et al. 2004. *Biosci Biotechnol Biochem* **68**:1315–1323.
11. Mandaokar A, et al. 2003. *Plant Mol Biol* **52**:775–786.
12. Wang Z, et al. 2005. *Plant Mol Biol* **58**:721–737.
13. Wang CS, et al. 1992. *Am J Bot* **79**:118–127.
14. Sheoran IS and Saini HS. 1998. *J Plant Biochem and Biotech* **7**:47–50.
15. Imin N, et al. 2001. *Proteomics* **1**:1149–1161.
16. Kerim T, et al. 2003. *Proteomics* **3**:738–751.
17. Kerim T, et al. 2003. *Funct Plant Biol* **30**:843–852.
18. Imin N, et al. 2004. *Proteomics* **4**:1873–1882.
19. Imin N, et al. 2006. *Mol Cell Proteomics* **5**:274–292.
20. Agrawal GK and Rakwal R. 2006. *Mass Spec Rev* **25**:1–53.
21. Saruyama H and Tanida M. 1995. *Plant Sci* **109**:105–113.

22. Wen L, et al. 2006. *Yi Chuan* **28**:311–316.
23. Gygi SP, et al. 1999. *Nat Biotech* **17**:994–999.
24. Ideker T, et al. 2001. *Science* **292**:929–934.
25. Kerk NM, et al. 2003. *Plant Physiol* **132**:27–35.
26. Casson S, et al. 2005. *Plant J* **42**:111–123.
27. Ai J, et al. 2006. *Proteomics* **6**:538–546.
28. Gozal YM, et al. 2006. *Methods Enzymol* **412**:77–93.

POLLEN PROTEOME

Sandra Noir

18.1 INTRODUCTION

The male gametophyte (or pollen) in higher plants is a key mediator of sexual reproduction. The last two decades have been marked by increasing efforts to decipher the genetic and molecular basis of pollen development and functions. Many of the biochemical events that occur in the vegetative cell of the pollen grain prior to anthesis (pollen maturation) are related to the preparedness of the gametophyte for the exigencies of germination and growth of the pollen tube in the pistil. In most plants, to achieve this considerable growth the pollen grain accumulates and stores large amounts of both mRNAs and proteins [1]. At present, many questions still remain. Which mRNAs are stored in mature pollen grains, and when are these mRNAs translated into proteins? Are there pre-synthesized proteins in mature pollen grains which are required for pollen germination and pollen tube elongation [1, 2]? If so, what are the proteins involved in signaling during pollen development or during fertilization [3]? Indeed, until recently, little biochemical and molecular genetic information was available about pollen functions. Recent studies have notably broken this bottleneck by performing transcriptomics analyses from pollen of the model plant *A. thaliana*

[4, 5]. Global gene expression analysis is in fact useful for selecting candidates for functional studies. Nevertheless, a gene transcript in eukaryotic cells usually gives rise to multiple protein isoforms, possibly with different functions, after alternative splicing and/or essential PTMs [6]. Therefore proteomic studies have been acknowledged as essential in unraveling the biological complexity of the pollen grain and for obtaining a better understanding of pollen functions.

18.2 BIBLIOGRAPHIC REVIEW: POLLEN DEVELOPMENT AND FUNCTION

Pollen—The Male Gametophyte

In flowering plants (angiosperms), development of the male gametophyte is a well-programmed and elaborate process. Male sporogenesis begins with the division of a diploid sporophytic cell, giving rise to tapetal initials and sporogenic cells. Figure 18.1 illustrates the subsequent events leading to the formation of a pollen grain. These sporogenic cells then undergo several mitoses to differentiate into pollen mother cells. Subsequently, each diploid pollen mother cell forms a callose-encased tetrad of haploid microspores by two consecutive meiotic divisions. The haploid cells of the tetrad are released as free microspores by the action of an enzyme (callase) produced by the tapetum layer of the anther [7]. Each uninucleate microspore then undergoes an asymmetric mitotic division forming a large vegetative cell and a smaller generative cell entirely enclosed within the vegetative cell (i.e., bicellular pollen stage). In nearly 70% of flowering plants (e.g., Solanaceae, Liliaceae), the pollen grain is shed from the anther at the bicellular pollen stage; the second mitotic division of the generative cell, which gives rise to two sperm cells (i.e., tricellular pollen stage), occurs while the pollen tube grows through the female pistil. In the remaining plant families (e.g., Cruciferae, Graminae), this second division occurs before anther dehiscence proceeds [8].

When the pollen is released from the anther, it exists as a free organism until it is carried by wind, insects, or other agents to the stigma of the pistil, where it starts a new phase of its life cycle. At this stage of development, if the pistil is compatible, the vegetative cell controls the germination of the mature pollen grain and growth of the pollen tube into the style. During this phase of development, there is an intimate interaction between the pollen tube and the cells and tissues of the pistil. Finally, sperm cell nuclei are delivered to the embryo sac in the ovule, where they participate in double fertilization [1].

The Mature Pollen Grain: An Attractive Biological Model System

During the last few decades, the increased interest in understanding pollen biology has been primarily related to its obligatory function in plant reproduction. However, the extremely reduced haploid male plant (male gametophyte) of flowering plants also constitutes a very attractive biological model system. The pollen corresponds to

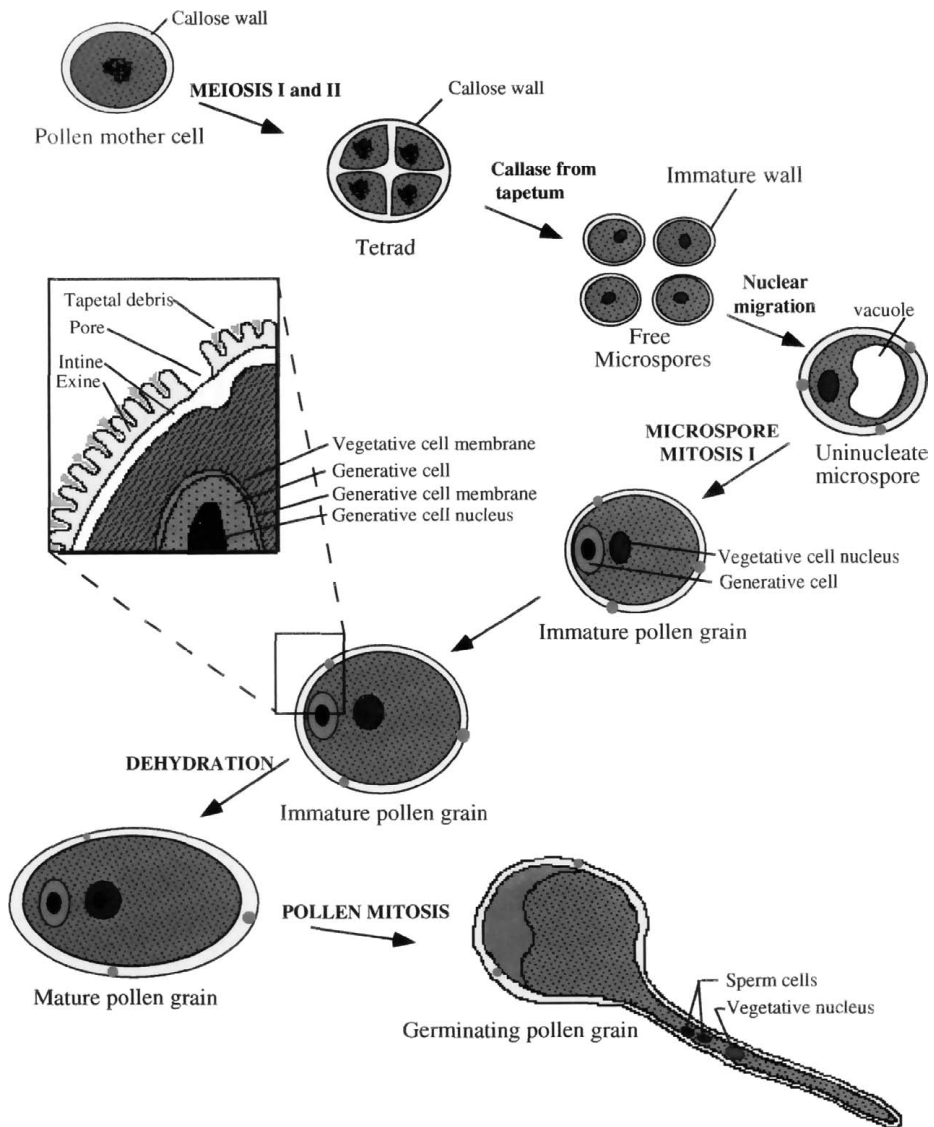


FIGURE 18.1. Schematic presentation of microsporogenesis typical of most angiosperms. (From McCormick [7], Copyright American Society of Plant Biologists.)

a homogeneous tissue type (only three cells) with a defined and rather small transcriptome and thus, very likely, proteome also. For instance, a total of 13,977 male gametophyte-expressed mRNAs have been identified from the different pollen stages of *A. thaliana* [9], which correspond to only around 40% of the full genome of the model plant. In addition, for a majority of monocots (e.g., maize, rice) and dicots (e.g., Solanaceous species), mature pollen is able to be easily collected at dehiscence

in bulk, either by shaking tassels/panicles into paper bags or by agitating flowers with a mechanical buzzer [10]. In the case of the quite small flowers of the *Arabidopsis* model plant, Johnson-Brousseau and McCormick [10] developed a vacuum system allowing the harvest of milligram amounts of mature pollen. The isolation of earlier stages of pollen grains (i.e., microspore and bicellular pollen) requires heavier procedures, since grains are embedded in the anthers. Nevertheless, this is a perfectly feasible possibility [9]. Consequently, in addition to its role in plant reproduction, the pollen grain has progressively become a very useful model for studying fundamental aspects of plant biology such as cell fate determination [11], cell–cell interactions [12], or cell polarity and tip growth [13, 14]. Due to those attractive considerations, pollen constitutes an ideal experimental system for analyses of key biological processes in higher plants.

18.3 METHODOLOGY AND EXPERIMENTAL RESULTS

In the context of pollen proteome analyses, main breakthroughs have been achieved using *O. sativa* and *A. thaliana*, monocot and dicot reference plant species, respectively. Indeed, the completion of the genome sequence of *Arabidopsis* [15] and *O. sativa* [16] generated considerable sequence database information and made these plants ideal for proteomics studies. The following section will focus on studies related to these two plants and mainly refer to the publications of Holmes-Davis et al. [3], Noir et al. [17], and Dai et al. [18].

Establishment of Mature Pollen Reference Maps

Until recently, only a few detailed proteomic studies have been carried out on the plant male gametophyte. The most comprehensive of these analyses were the proteome reference maps that have been generated from rice anthers to develop information about proteins present at different stages of pollen development [19, 20]. In the latter study [20], 2-DGE and MS approaches have been used to analyze the protein complement of rice anthers, and the authors were able to identify approximately 150 protein spots that varied among different stages of anther development (for further details, please refer to Chapter 17 of this book). In addition to the rice anther studies, a proteomic analysis of the *Arabidopsis* pollen coat has been performed to identify components needed for efficient pollination [21]. Following cyclohexane extraction (which improves extraction of highly hydrophobic proteins) [22], 12 proteins were identified, nine of which were oleosins (glycine-rich proteins) or lipases. These proteins along with lipids are thought to support the initial signaling interactions between the pollen and pistil [21].

Finally, owing to the increasing level of data available on *Arabidopsis* pollen transcriptional profiling [4, 5, 9, 23], proteome analyses have been performed from mature pollen of *Arabidopsis* [3, 17] and also *O. sativa* [18]. In the analysis of Holmes-Davis and co-workers a multistep procedure was used to extract proteins from mature pollen of *A. thaliana* to obtain polypeptides with different solubility properties; salt-soluble and salt-insoluble proteins, as well as SDS-extracted proteins, were distinguished.

These extracts were analyzed by 2-DGE, and from the gels of these three extracts a combined total of 225 spots were excised. These were chosen to ensure a distribution of high- and low-abundance proteins and to include proteins from a wide range of MWs and pIs. By using ESI–MS/MS, 179 peptides were identified, representing 135 distinct proteins. In the study of Noir and co-workers, total proteins were extracted from mature pollen of *A. thaliana* using a CHAPS-detergent buffer that improves extraction of PM-associated proteins supposedly implicated in pollen development and pollen tube growth. Protein samples were subjected to 2-DGE, and 121 different proteins were identified from 145 spots reproducibly resolved and analyzed by MALDI–TOF–MS and/or LC–MS/MS techniques. Despite the property of the SDS buffer to extract a wide variety of proteins with different solubilities, only 43 identical polypeptides were identified in these two studies.

In the analysis of Dai et al. [18], a two-step procedure was applied to fractionate *O. sativa* mature pollen proteins into pollen-released and pollen-interior proteins. In addition, using diethyl ether, a pollen coat protein-enriched fraction was prepared. In sum, by using MALDI–TOF–MS and/or ESI–MS/MS techniques, the identity of 545 peptides was determined representing 322 unique proteins. Comparing these data with the two analyses performed on *Arabidopsis* mature pollen and considering the differences in annotation between *A. thaliana* and *O. sativa* species, 53 and 44 equivalent proteins were identified in the reference maps of Holmes-Davis et al. [3], and Noir et al. [17], respectively. Comparing the three analyses, 22 proteins exhibiting similar identifications were found to be common. Due to specific treatments required to analyze the pollen coat proteome, none of the *Arabidopsis* pollen coat proteins described in the analysis of Mayfield et al. [21] were identified in the two above-mentioned *Arabidopsis* mature pollen proteome maps. Likewise, despite the analysis of a rice pollen coat-related protein fraction from which 35 proteins were identified [18], none of them (i.e., rice homologs) were described by Mayfield et al. [21].

Isoforms

In each of the three analyses, a significant proportion of proteins—17% [3], 30% [17], and 23% [18]—were identified multiple times. Generally, proteins can be assigned to two classes of isoforms: (a) polypeptide variants encoded by the same gene resulting from PTMs or alternative splicing and (b) sequence-related protein isoforms encoded by distinct genes (paralogs). Both kinds of isoforms were detected in mature pollen proteome analyses. It also appears that PTMs such as phosphorylation, glycosylation, ubiquitination, or acetylation could play a significant role during pollen development by diversifying the function of a given protein.

Functional Categories of Mature Pollen Proteins

From the three mature pollen reference maps [3, 17, 18], the authors determined the predicted functional categories assigned to the identified proteins. Despite differences in the category classification chosen by the authors, similar functional classes were revealed as being predominant in the three analyses. These major functional categories (Table 18.1) are discussed below.

TABLE 18.1. Functional Classification Proportions (%) of the Proteins Identified from Mature Pollen in the Analyses of Holmes-Davis et al. [3], Noir et al. [17], and Dai et al. [18]^a

	<i>Arabidopsis thaliana</i>		<i>Oryza sativa</i>
	Holmes-Davis et al. [3]	Noir et al. [17]	Dai et al. [18]
Metabolism	20	20	8 ^b
Energy generation	17	21	25 ^c
Protein processing ^d	18	10	14
Cytoskeleton dynamics and cell wall remodeling ^e	12	9	14
Cellular transport and signaling ^f	17	4	16
Defense and stress response	6	4	4
Cell cycle and DNA processing	2	1	—
Transcription-related	—	—	2
Others	—	2	5
Unclassified	8	27	11
<i>De novo</i> proteins	—	—	2

^aTo facilitate the comparison, functional categories have been modified in this table.
^bNucleotide acid, amino acid, and lipid metabolisms.
^cCarbohydrate and energy metabolism.
^dProtein synthesis, and protein destination and storage [3]; protein synthesis and protein fate [17]; protein synthesis, assembly, and degradation [18].
^eCell structure [3]; biogenesis of cellular compounds [17]; cytoskeleton dynamics, and wall remodeling and metabolism [18].
^fIntracellular traffic, signal transduction, and transport [3]; cellular transport and cellular signaling [17]; signal transduction and ion transport [18].

Metabolism and Energy Generation. As mentioned previously, the role of the pollen grain in plant reproduction involves the production of a tube in order to deliver the sperm cells to the embryo sac. It therefore seems logical that the pollen grain stores proteins required for pollen tube germination and early growth, as well as proteins for energy metabolism [1]. In each of the three analyses, a high proportion of polypeptides are involved in general metabolism (Table 18.1). Similarly, important percentages of proteins (at least 17%, Table 18.1) required for energy generation were identified (e.g., ATP-synthase, enolase, or GAPDH).

Protein Processing. Interestingly, more than 10% of the identified proteins were assigned to protein processing functions (Table 18.1). It was suggested by Honys and Twell [5] that mature pollen is charged with a preformed translational apparatus enabling rapid activation upon hydration and germination. However, whether this is true and, if so, which components of the translation machinery are presynthesized in mature pollen was so far still unclear. In the three analyses of reference, major proteins

implicated in protein synthesis such as ribosomal proteins or translational initiation and EFs have been identified. Furthermore, two putative inhibitors of translation have been described from the rice proteome map [18]. Other proteins more related to protein fate functions, such as chaperone/chaperonin, polyubiquitin or cyclophilin, were identified. All together, these data provide molecular evidence at the protein level that mature pollen in flowering plants presynthesizes the necessary components of the translation machinery to initiate protein synthesis and destination upon entry into germination [18].

Cytoskeleton Dynamics and Cell-Wall Remodeling. Germination of pollen grains require polarization by cytoplasmic and cytoskeletal reorganization; after that pollen tube growth is restricted to the tip region, which demands a continuous deposition of new cell wall and PM [1]. Considering the different annotations in the three analyses previously mentioned, a significant proportion of identified proteins were classified as related to cytoskeleton dynamics and cell-wall remodeling functions (>9%, Table 18.1). Different isoforms of actin, tubulin, and also actin-binding proteins, namely profilin and actin-depolymerizing factors, which regulate the dynamics of the actin cytoskeleton in the growing pollen tube [24], have been identified in mature pollen proteomes. A typical function of these cytoskeletal proteins is their involvement in the deposition of cell-wall components such as pectin, cellulose, and callose [25]. Likewise, proteins required for pollen wall dynamics were identified. *Arabidopsis* and rice mature pollen proteomes include both proteins involved in degradation of cell-wall components such as cellulase, polygalacturonase, exopolygalacturonase, pectinesterase, and proteins involved in active wall synthesis (e.g., cellulose synthase, reversibly glycosylated polypeptide). The detection of these different cytoskeleton and cell-wall components presynthesized in mature pollen grains suggests complex machinery ready to mediate reorganization leading to pollen germination and fast tube elongation and orientation.

Cellular Transport and Signaling. In view of the previously mentioned physiological and metabolic events that occur in the mature pollen grain, it is understandable that signaling molecules were detected in *Arabidopsis* and rice mature pollen proteomes. Indeed, considering the different annotations, this function was assigned to 4% to ~15% of the identified proteins (Table 18.1). Several of the proteins involved in signal transduction are either (a) Ca^{2+} -sensors (e.g., CaM-like protein, calreticulin), kinases, and proteins mediating phosphorylation/dephosphorylation-activity or (b) proteins implicated in GTP-mediated signaling. Most of the identified proteins implicated in transport are anion-channel proteins, H^{+} -transporting ATPases, or vacuolar ATP synthases.

Defense and Stress Response. Unexpectedly, a significant portion of stress-related proteins were also identified in the three pollen proteome analyses (~5%, Table 18.1). The identification of proteins such as HSPs, POX, catalases, peroxiredoxins, or GSTs suggests a global capability of the mature pollen grain to protect itself from biotic and abiotic stresses. As mentioned by Dai et al. [18], the pollen, as

a highly compact tricellular organism, must have acquired the ability during evolution to deal with (a) extracellular stresses after release from the anther and (b) intracellular stresses caused by the active metabolism of germinating pollen and its interaction with cells of the stigma and the style.

Unknown Function. Finally, in the three analyses a significant number (>8%) of the identified polypeptides are of unknown function (Table 18.1). By analogy, this set of proteins may hint at as-yet-unidentified cellular processes, some of which might be specific for the male gametophyte.

Comparison of Proteomics and Transcriptomics Analyses of *Arabidopsis* Pollen

With the development of comprehensive *Arabidopsis* gene chips, large-scale analyses of the *Arabidopsis* pollen transcriptome have been carried out during the last four years [4, 5, 9, 23]. Therefore this section will focus only on analyses performed on this model plant.

In terms of functional classification, Figure 18.2 illustrates the comparison between mature pollen proteomic data [3, 17] and transcriptomic data [23]. Similar proportions related to general metabolism (20%), protein processing (14%), and stress-related (5%) functions are observed in the proteomic data compared to the transcriptomic data (21%, 10%, and 2%, respectively). Polypeptides devoted to transport and signaling (10%) functions appear underrepresented in proteome analyses—and in even larger proportion in the analysis of Noir et al. [17]—compared to gene expression data (27%). Indeed, proteins implicated in these functions are often either of low-abundance and/or integral membrane proteins that are difficult to detect in 2-DGE-based approaches. Low abundance may also account for a similar disproportion regarding transcription-related regulators (0% in both proteome analyses versus 17% in transcriptome). Gene products exhibiting functions in both energy metabolism (~19%) and cytoskeleton and cell-wall metabolism functions (~11%) appear to be more highly represented in the proteome compared to the transcriptome (7% and 6%, respectively). This may indicate that mature pollen is well-prepared with preexisting metabolic enzymes and cellular components for rapid energy conversion upon germination and pollen tube expansion. It must be noted here that authors of the transcriptomic analysis do not mention any genes of unclassified function.

Expression data of genes corresponding to proteins identified in *Arabidopsis* mature pollen proteome studies were examined by expression analysis on Affymetrix ATH1 microarray and EST representation [3], and data from Noir et al. [17] were reinvestigated by using updated microarray data deposited at the Genevestigator server (October 2006 [26]). These analyses revealed 13 (At1g26320, At2g16730, At3g01250, At3g05930, At3g08560, At4g24640, At4g25590, At4g29340, At5g25880, At5g52360, At5g59370, At5g61720, At5g62750) and 14 (At1g49490, At1g64980, At1g67290, At1g69940, At1g78300, At3g05930, At3g07850, At3g46520, At4g24640, At5g04180, At5g44340, At5g52360, At5g59370, At5g61720) distinct genes, respectively, specifically expressed in the pollen organ. Four of these pollen-specific genes are common

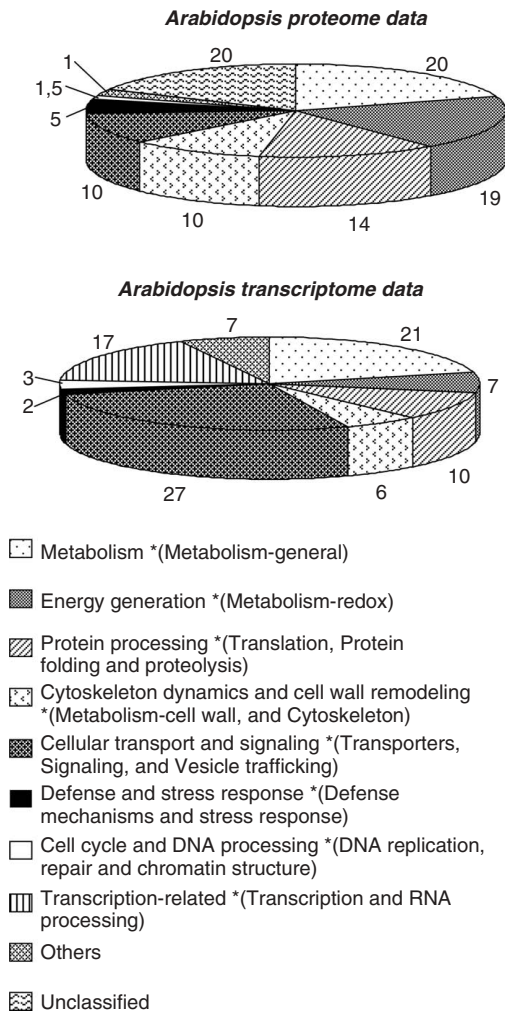


FIGURE 18.2. Comparison of the functional classification established from *Arabidopsis* mature pollen proteomic and transcriptomic analyses. Proteomic data correspond to the average ratio of Holmes et al. [3] and Noir et al. [17] results (Table 18.1). Transcriptomic data relate to the study of Pina et al. [23] and are based on the analysis of pollen expressed-genes using the Affymetrix ATH1 GeneChip microarray. The function annotation has been modified as indicated (*) to facilitate the comparison.

to the two analyses (At3g05930, At4g24640, At5g59370, and At5g61720). Most of the proteins identified as pollen-specific are related to cytoskeleton and cell-wall metabolism functions. These include germin-like protein (At3g05930), exopolysaccharuronase (At3g07850), pectinesterase (At1g69940), and pectin methylesterase inhibitor

protein (At4g24640), actin (At5g59370, At3g46520) and tubulin (At5g44340) isoforms, actin-depolymerizing factors (At4g25590, At5g52360), and extensin-like protein (At1g49490). Interestingly, four of these pollen-specific proteins (At1g64980, At3g01250, At5g61720, At5g62750) are so far unclassified, suggesting that these polypeptides may represent new protein functions.

In addition, each protein identified in both analyses [3, 17] has been compared with its respective transcriptomic data; transcripts of 10 and nine distinct proteins, respectively, were not detectable in transcriptomics analyses. None of these 19 proteins are similar, and none is reported as being pollen-specific. For proteins predicted to be secreted (At2g17420, At2g39960, and At3g02230), an explanation of their presence might be that they were deposited by the tapetum [3]. Alternatively, for any of the other candidates, respective transcripts might either be absent in mature pollen or present at very low levels. This observation implies that low transcript abundance may suffice to ensure accumulation of a significant proportion of protein in some instances (e.g., for proteins with a low turnover rate). Alternatively, the respective polypeptides might have been translated at an earlier stage of pollen development (e.g., in the bicellular stage) and still persist in mature pollen. In a more detailed approach, Holmes-Davis et al. [3] compared the protein levels of 77 candidates with their corresponding mRNA signal. Despite some exceptions, the authors generally observed an inverse relationship: mRNAs that were abundant corresponded to lower-abundance proteins while mRNAs that were less abundant corresponded to some of the more abundant proteins. These latter proteins were primarily involved in energy generation, while the former were mainly involved in cytoskeleton and cell-wall metabolism. According to the authors, it seems reasonable that the mature pollen grain stores proteins that it will need for the immediate generation of energy to initiate germination and pollen tube growth. It appears also plausible that mRNAs encoding proteins necessary for cell structure are ready for immediate translation upon germination of the pollen grain. In any case, these findings further underline the importance of proteome studies as valuable supplements to transcriptomics analyses.

An Example of Application: Pollen Allergen Analysis

Inhalative allergies are elicited predominantly by pollen of various plant species. A large number of grass, weed, and tree species shed their pollen in high concentrations during the pollen season, leading to allergic symptoms such as hay fever, rhinitis, and even bronchial asthma [27]. The last two decades were marked by large advances in the characterization of pollen-derived allergenic proteins.

Recently, proteome and immunochemical approaches have been extensively used to analyze the allergen content of pollen grains [28–30]. After 1-DGE or 2-DGE, pollen extracts were transferred from gels to nitrocellulose membranes and allergens revealed by immunodetection using serum (concentrated in allergen-specific immunoglobulin E-IgE) from given allergic patients. After gel-excision and appropriate digestion, IgE-reactive proteins were subsequently identified by MALDI-TOF-MS or MS/MS techniques. To improve the detection of low-abundance IgE-reactive proteins, pollen protein extracts can be chemically fractionated [31] or depleted by using an affinity column [30].

So far, most of the described pollen allergens are water-soluble proteins or glycoproteins of MWs from 10 to 70 kDa [32]. They have been identified in diverse plant species and the more commonly reported correspond to varied protein types. For example, major pollen allergens—such as the β -expansins, glycoproteins that mediate cell-wall extension in plants [33], or the pectate lyases, which are effective cell-wall-degrading enzymes [34]—are implicated in cell-wall dynamics. Likewise, profilins, which play an active role in the regulation of actin polymerization [35], or members of PR protein families (namely, TLP) have been reported as allergens in many plants [36]. Interestingly, a global study, based on the sequence analysis of 157 reported pollen allergens, revealed that these polypeptides belong to only 29 distinct protein families (among a total of 7868 protein families in the Pfam database) [37]. In this analysis, Radauer and Breiteneder [37] demonstrate that expansins, profilins, and calcium-binding proteins (EF-hand proteins) constitute the major pollen allergen families.

Nevertheless, due to numerous isoforms of pollen allergenic proteins and great variation in patient susceptibility, the repertoire of pollen allergens is far from being exhaustive. Given the clinical value of this research area, proteomic combined with immunoserological approaches will still be necessary to identify new pollen allergens, as well as to predict the allergenic potential of novel pollen proteins.

18.4 CONCLUSIONS

The improvements in techniques to collect pollen in adequate amounts and of proteome approaches facilitating polypeptide identification using a minimum of material have allowed researchers to engage in large-scale proteomic analyses of mature pollen. These analyses have been performed on the two reference plants *Arabidopsis* and *O. sativa*, since bioinformatic data from both of these species enable convenient protein identifications. So far, a total of 438 distinct polypeptides have been identified from mature pollen [3, 17, 18]. The majority (more than 60%) of these proteins are allocated to five main functional categories: metabolism, energy generation, protein processing, cytoskeleton dynamics and cell-wall remodeling, and cellular transport and signaling. These data confirm at the protein level that during maturation, pollen grains in flowering plants accumulate a large set of proteins dedicated to further biological processes. In complement to previous transcriptional analyses of *Arabidopsis* pollen [4, 5, 9, 23], these proteomic reports show that mature pollen stores, at least in part, the protein machinery that is needed for germination, as soon as the required environmental conditions are satisfied, and for further pollen tube elongation and guidance. Despite the presently incomplete description of the mature pollen proteome, a significant list of proteins is now available and can contribute to the understanding of the biochemistry of the pollen grain. A further consideration will be to shed light on the temporal and spatial coordination of this protein network in the pollen grain itself and then during its intimate interaction with the cells and tissues of the pistil. Furthermore, a relevant proportion of polypeptides identified in mature pollen is still of unclassified function. These proteins provide a basis for the use of reverse genetics

to identify novel biological functions in plants by taking advantage of the reduced complexity of the pollen system.

18.5 FIVE-YEAR VIEWPOINT

The generation of pollen proteome data is still in its infancy. More remains to be done! Up to now, these proteomics resources have been restricted to one stage of pollen development. Initially, further experimentation will be necessary to complete the repertoire of the mature pollen proteome. In this perspective, research efforts should be applied to unravel the proteome of the male gamete (i.e., sperm cells). The highly specialized functions of these sperm cells in double fertilization are expected to be controlled by a unique set of proteins [38]. Then, analyses of early stages of pollen maturation (i.e., microspore and bicellular pollen) will also require additional investigations and will allow the identification of stage-specific pollen proteins. In this respect, it would be interesting to use well-characterized mutant pollen to better understand such phenotypes as male sterility. Likewise, functional pollen proteomic is an emerging area of research, and so far these studies are nonexistent, to my knowledge, with regard to *Arabidopsis* or rice pollen. It would be interesting to reveal differential protein expression profiles between distinct developmental steps, such as mature pollen, germinated pollen, and pollen with elongated tube. These analyses have been initiated only from pollen of gymnosperms (seed plants) to characterize proteomes of pollen tubes in *P. strobus* [39] and *P. meyeri* [40]. Similarly and in the context of global climatic changes, the idea of comparing pollen proteomes under abiotic stress (e.g., heat or cold stress, high concentration of carbonic acid) is very attractive. Given the power of the proteomic approach in plant biology studies, I hope that these considerations will be of interest to some research teams in the further coming years.

ACKNOWLEDGMENTS

I acknowledge Ralph Panstruga for giving me the opportunity to write this chapter and for his interesting comments, as well as Cristina Micali and Matt Humphry for critical reading of the manuscript. I would like as well to highlight the works of Holmes-Davis et al. [3], and Dai et al. [18], which allowed me to give more relevance to this chapter.

REFERENCES

1. Mascarenhas JP. 1993. *Plant Cell* 5:1303–1314.
2. Twell D. 1994. The diversity and regulation of gene expression in the pathway of male gametophyte development. In: *Molecular and Cellular Aspects of Plant Reproduction* (Scott RJ and Stead AD, eds.), Seminar series, 55 (Society for Experimental Biology), Cambridge University Press, Cambridge, pp. 83–135.

3. Holmes-Davis R, et al. 2005. *Proteomics* **5**:4864–4884.
4. Becker JD, et al. 2003. *Plant Physiol* **133**:713–725.
5. Honys D and Twell D. 2003. *Plant Physiol* **132**:640–652.
6. Lockhart DJ and Winzeler, EA. 2000. *Nature* **405**:827–836.
7. McCormick S. 1993. *Plant Cell* **5**:1265–1275.
8. Brewbaker JL. 1967. *Am J Bot* **54**:1069–1083.
9. Honys D and Twell D. 2004. *Genome Biol* **5**:R85.
10. Johnson-Brousseau SA and McCormick S. 2004. *Plant J* **39**:761–775.
11. Twell D, Park SK, and Lalanne E. 1998. *Trends Plant Sci* **3**:305–310.
12. Kachroo A, Nasrallah ME, and Nasrallah JB. 2002. *Plant Cell* **14**:S227–S238.
13. Hepler K, Vidali L, and Cheung AY. 2001. *Annu Rev Cell Dev Biol* **17**:159–187.
14. Feijó JA, et al. 2004. *Curr Opin Plant Biol* **7**:589–598.
15. *Arabidopsis* Genome Initiative. 2000. *Nature* **408**:796–815.
16. International Rice Genome Sequencing Project. 2005. *Nature* **436**:793–800.
17. Noir S, et al. 2005. *Biochem Biophys Res Commun* **337**:1257–1266.
18. Dai S, et al. 2006. *Proteomics* **6**:2504–2529.
19. Imin N, et al. 2001. *Proteomics* **1**:1149–1161.
20. Kerim T, et al. 2003. *Proteomics* **3**:738–751.
21. Mayfield JA, et al. 2001. *Science* **292**:2482–2485.
22. Mayfield JA and Preuss D. 2000. *Nature Cell Biol* **2**:128–130.
23. Pina C, et al. 2005. *Plant Physiol* **138**:744–756.
24. Chen CY, et al. 2002. *Plant Cell* **14**:2175–2190.
25. Cai G, et al. 2005. *Plant Cell Physiol* **46**:563–578.
26. Zimmermann P, et al. 2004. *Plant Physiol* **136**:2621–2632.
27. Shivanna KR. 2003. Pollen morphology and aeropalynology. In: *Pollen Biology and Technology* (Shivanna KR, ed.), Science Publishers Inc., Enfield, NH, pp. 26–44.
28. Kerim T, et al. 2003. *Funct Plant Biol* **30**:843–852.
29. Corti V, et al. 2005. *Proteomics* **5**:729–736.
30. Kao SH, et al. 2005. *Proteomics* **5**:3805–3813.
31. Napoli A, et al. 2006. *Anal Chem* **78**:3434–3443.
32. Puc M. 2003. *Ann Agric Environ Med* **10**:143–149.
33. Li L-C and Cosgrove DJ. 2001. *Eur J Biochem* **268**:4217–4226.
34. Wing RA, et al. 1990. *Plant Mol Biol* **14**:17–28.
35. Valenta R, et al. 1991. *Science* **253**:557–560.
36. Breiteneder H. 2004. *Allergy* **59**:479–481.
37. Radauer C and Breiteneder H. 2005. *J Allergy Clin Immunol* **117**:141–147.
38. Engel ML, et al. 2003. *Plant J* **34**:697–707.
39. Fernando D. 2005. *Proteomics* **5**:4917–4926.
40. Chen Y, et al. 2006. *Plant J* **47**:174–195.

MICROTUBULE-BINDING PROTEINS

Lori A. Vickerman and Douglas G. Muench

19.1 INTRODUCTION

Microtubules (MTs) participate in a number of essential processes in plant cells, including cell division, cell morphogenesis, and cell-wall formation. The organization and dynamics of MTs must be tightly controlled in order to efficiently carry out these processes. This depends largely on the activity of MT-associated proteins (MAPs). In the past decade, several plant MAPs have been identified, some of which have no apparent equivalent in animal, protist, or fungal systems. The specific roles in the organization and dynamics of the four plant MT arrays is becoming clearer for many of the plant MAPs. In addition to the structural MAPs, numerous other proteins bind to MTs in plant cells. These proteins are predicted to use MTs as a matrix to transiently position themselves in close proximity with other proteins or to specific regions of the cell, or they may bind to MTs as a means to regulate their own activity. It is estimated that there are hundreds of proteins that interact with MTs at one stage or another during the cell cycle. The importance of MTs in plant growth and development has made the identification and characterization of these MT-binding proteins a research focus for numerous laboratories. Research progress in this area has been enhanced

by technological advances in MS, live cell imaging, and the availability of genomics related resources.

19.2 BRIEF BIBLIOGRAPHIC REVIEW

MTs are dynamic structures formed by the association of linear protofilaments composed of polymers of repeated α - and β -tubulin subunits. This design results in an inherent polarity: β -Tubulin is exposed on the faster-growing plus end of MTs, and α -tubulin is exposed on the less dynamic and slower growing minus end. In plant cells, there are four distinct MT arrays, three of which—the pre-prophase band, the phragmoplast, and the metaphase spindle—are involved in the process of cell division [1]. The pre-prophase band and the phragmoplast are unique to plants, and they have essential roles in the formation and orientation of the cell plate. The spindle has an evolutionarily conserved role in both plant and animal cells in the segregation of chromosomes during cell division. The fourth array, the interphase cortical array, forms a highly organized network of MTs that lines the PM. Cellulose microfibrils, the strengthening component of the plant cell wall, tend to be aligned in parallel with the cortical MTs. Mounting evidence suggests that MTs regulate the directional movement of the cellulose synthesizing machinery. Although the precise roles and mechanisms for the MT–cellulose synthase relationship are still under debate, the identification of new mutants and advances in live cell imaging have provided clues toward understanding this process [2, 3]. Additional roles for the plant cortical MT network have emerged. For instance, the close proximity of this network at the outer boundary of the cytoplasm places it in an ideal position to perceive and transmit environmental signals to the cell. Indeed, abiotic and biotic stress, as well as hormones, can trigger dramatic changes in MT organization that may be important in releasing MT-bound signaling molecules into the cell [discussed in references 3 and 4].

MAPs have classically been defined as proteins that alter the function or behavior of MTs, and that can be co-purified with MTs *in vitro* after cycles of MT assembly and disassembly [5]. MAPs regulate MT organization and dynamics by cross-bridging, stabilizing, and severing MTs, or linking MTs to actin filaments or membranes. Often placed into this group of classical MAPs are motor-related proteins that are involved in MT-mediated movements and interactions between MTs and other cellular structures. The term “MAP” has also been used to loosely describe numerous other proteins that bind to MTs but that do not function to regulate the dynamics or organization of MTs. These nonclassical MAPs or “MT-interacting proteins” [4, 6] probably bind transiently to MTs to perform several possible functions. For example, these protein–MT interactions may serve to regulate the activity of a protein, or they may be important in localizing or concentrating a protein in a specific region of the cell. These interactions may also be important in the assembly of protein complexes, perhaps to enhance processes such as translation or to facilitate metabolic channeling. We have used the term “MT-binding protein” throughout this chapter when referring collectively to classical and nonclassical MAPs.

The organization of MTs into sheets and bundles provides a large surface area for these dynamic interactions between MTs and MT-binding proteins [7]. The large

number of confirmed and predicted MT-binding proteins indicates that the MT surface is extremely congested. This suggests that the binding of these proteins to MTs is under tight temporal and spatial control and is influenced by cell type, physiological state, and external influences. This chapter focuses on approaches used to characterize the plant MT-binding proteome and discusses examples of MT-binding proteins that have been characterized in detail. Two recent, large-scale biochemical studies aimed at identifying MT-binding proteins are also discussed, as are the possible functions of “nonclassical” MAP interactions with the cytoskeleton.

19.3 STRATEGIES AND METHODOLOGIES

Informatics Approaches Using Plant Genome Sequence Information

Data generated from plant genome sequencing projects have assisted in the identification of a number of homologs of known MT-binding proteins [8, 9]. This has been extremely useful in identifying plant MAPs that have not surfaced using other approaches. Sequence database searches have also provided information on the level of redundancy of genes encoding MT-binding proteins in plants. This has led to more detailed analysis of tissue- and cell-specific expression, and the MT array-specific binding of individual MAP family members.

Mutant Screens

Forward genetic approaches have contributed significantly to the identification of several homologous or novel plant MT-binding proteins. Mutant populations produced by chemical mutagenesis and T-DNA/transposable element insertions have been used in screens for MT-binding proteins. The important role of MTs in cell plate formation and cellulose microfibril deposition have made genetic screens for cell shape and abnormal growth phenotypes useful in identifying structural MAPs. For example, mutations in plant MAPs often result in organ and stem twisting, as well as root swelling phenotypes [10–14]. In addition, a temperature-sensitive screen based on immunofluorescence imaging of MTs was successful in identifying a mutant impaired in MT organization [15].

Biochemical Purification Approaches

Co-Assembly and Co-Purification of MT-Binding Proteins Using Endogenous and Neuronal MTs. The ability to assemble tubulin dimers into MTs *in vitro* has been the basis for protocols that involve co-assembly or co-purification of MT-binding proteins with MTs. Protocols of this type have been typically modeled after those used for animal cells [16]. Preformed MTs, or MTs assembled directly from endogenous tubulin in plant cell extracts, are used as a matrix onto which MT-binding proteins are bound *in vitro*. MTs are then pelleted by centrifugation, and the associated proteins are released by high-ionic-strength treatment. Owing to the

low concentration of tubulin in most types of plant cells, mammalian neuronal tubulin is often used as the source of MTs in many plant MAP purification schemes. Brain tissue is rich in MTs, and standard protocols are available for tubulin purification from this tissue [17]. However, several reports have described the purification of MT-binding proteins using endogenous plant MTs assembled directly in cell extracts in the presence of the MT-stabilizing agent taxol [18–21]. A strategy to further enrich MT-binding proteins using this approach involves adding cycles of MT depolymerization and repolymerization to the pellet fraction [18]. Korolev et al. [22] demonstrated that the line of cell suspension used was essential for successful MT-binding protein purification with endogenous plant MTs. They concluded that small, dense cells of an *Arabidopsis* Landsberg cell line proved more effective for this procedure than the larger, vacuolated cells common to cell lines from other genotypes. Readers are directed elsewhere for a detailed description of methods used to purify MT-binding proteins via co-assembly of endogenous plant MTs [18, 20, 22] or by co-purification with exogenous neuronal MTs [23, 24].

MT- and Tubulin-Affinity Chromatography (MAC and TAC). MAC and TAC have proven effective as approaches for purifying authentic MT-binding proteins from a wide range of species. Protocols for column preparation and chromatography were based on those established for the isolation of *Drosophila* MT-binding proteins [25, 26]. Taxol-stabilized MTs or native tubulin dimers are covalently coupled to a chromatography matrix that often consists of activated agarose beads. A BSA column serves as a useful control column that can identify nonspecific interactions, since the overall pI of BSA is similar to that of tubulin. Clarified cell extracts are passed slowly over the column matrices, followed by column washing and then elution of bound proteins with solutions containing high concentrations of salt.

Interestingly, profiles of plant proteins that eluted from MAC and TAC columns were qualitatively similar when analyzed by SDS-PAGE, regardless of whether plant or neuronal tubulin was used as the source of tubulin [27, 28] (Figure 19.1A). Approximately 30–50 proteins are visible in these fractions when resolved by 1D SDS-PAGE slab gels and silver staining (Figure 19.1A). However, the complexity of proteins is more readily visible using 2-DGE and CBB [4, 28] or silver staining procedures (Chuong and Muench, unpublished observations). Indeed, well in excess of 100 proteins are clearly visible on 2D gels (Figure 19.1B), and numerous, more faintly staining protein spots are also present. We found that resolving these affinity chromatography purified proteins using 2-DGE required the use of NEPHGE techniques [29]. MT-binding proteins generally have a basic pI [28, 30], causing them to precipitate using the standard first-dimensional gel systems. However, this problem was overcome using NEPHGE. Here, the protein samples are loaded in tube gels that are immersed in acidic running buffer, thereby limiting the precipitation of this basic protein fraction [4, 28, 31].

The similarity between the SDS-PAGE profiles of proteins eluted from MT and TAC columns demonstrates that many proteins bind nearly as effectively to tubulin dimers as they do to MTs. Indeed, MT pelleting assays using tubulin-binding protein fractions indicated that a majority of the proteins in this fraction also bind

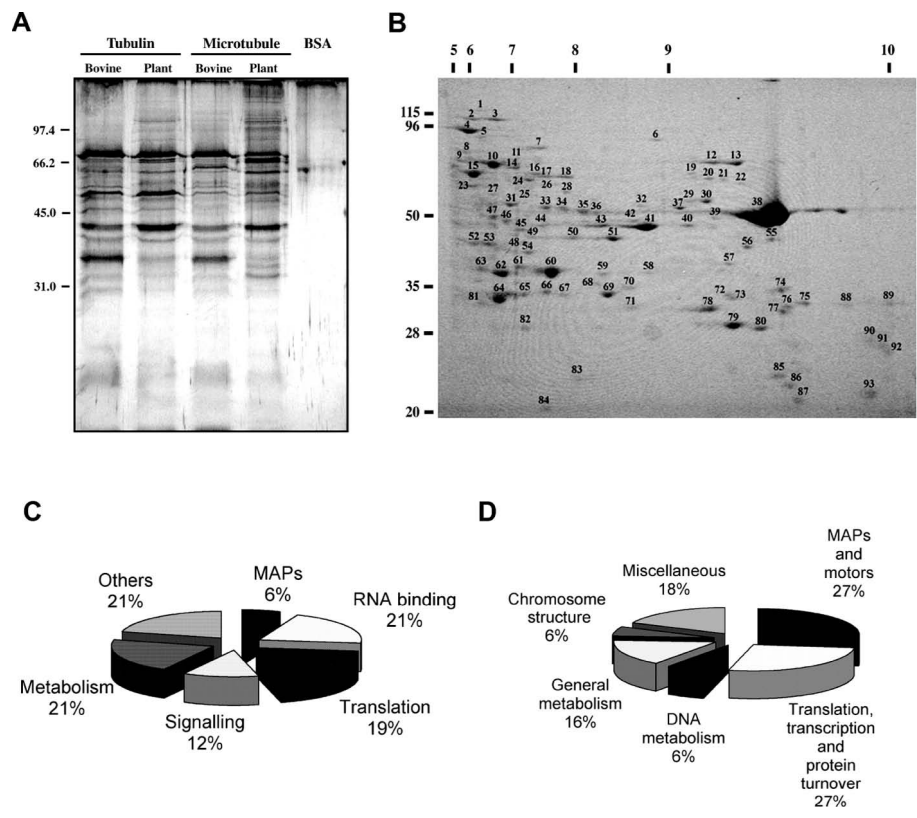


FIGURE 19.1. Protein profiles of MT- and tubulin-affinity chromatography-purified protein fractions, along with the class representations of MT-binding proteins from proteomic studies. **(A)** Silver-stained SDS-PAGE profiles of rice endosperm proteins that eluted from MT- or tubulin-affinity chromatography columns prepared using plant or bovine tubulin. Fractions were loaded on the columns in the presence of 50 mM KCl, washed, and eluted with 500 mM KCl. A BSA affinity column served as the control. (From reference 28 with permission from the publishers.) **(B)** CBB-stained 2D NEPHGE gel of an *Arabidopsis* cell suspension protein fraction eluted from a bovine tubulin-affinity column. The approximate pH value at specific locations in the first dimension gel is indicated at the top of the gel. Molecular mass standards are listed on the left of the gel. Candidate spots selected for LC-MS/MS analysis are numbered. (From reference 4 with permission from the publishers.) **(C)** Classes of proteins identified in a large-scale tubulin-affinity chromatography study. (From reference 4 with permission from the publishers.) **(D)** Classes of proteins identified in a large-scale MT co-assembly study [22]. Several tubulin isotypes were identified in this study, but were not included in this figure.

to MTs *in vitro* [4, 32]. TAC has been used to identify several authentic plant and animal MT-binding proteins [27, 28, 32, 33]. However, there are examples of structural MAPs that do not bind to tubulin dimers [34] and thus would not be purified using TAC. Close inspection of SDS-PAGE protein profiles showed that subtle differences are recognizable between the fractions purified by MAC versus TAC techniques

(Figure 19.1A). However, the high yield of tubulin in brain tissue extracts and the stability of TAC columns have made neuronal TAC an appealing approach for our laboratory.

19.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Most of the proteins that have been shown to bind MTs in plant cells are classical MAPs that have roles in regulating the dynamics and organization of MTs. Others fall into the “nonclassical MAP” category. In Table 19.1 we have listed examples of plant-specific proteins that have authentic MT-binding activity *in situ*, as well as proteins that are common to eukaryotes but have been shown to localize to MTs in plant cells only. Numerous other predicted MT-binding proteins have been identified using biochemical purification techniques or using genome database search engines, but have not yet been shown to localize to MTs *in situ*. In this section, we discuss the identity and roles of several of these confirmed and predicted MT-binding proteins. For more complete lists, readers are directed to reviews and research publications that provide detailed lists of these proteins [4, 8, 9, 22, 35].

MT-Binding Proteins Identified Using Informatics Approaches

Genome sequencing projects have provided a wealth of information on conserved genes and gene families that code for MT-binding proteins in plants. This information

TABLE 19.1. Proteins that Demonstrate an Authentic and Unique MT-Binding Activity in Plant Cells

Protein	MT Array	Known or Predicted Function	References
MAP70	All arrays		22
AIR9	Cortical, PPB	Cell plate maturation	57
p190	Phragmoplast	MT–actin interactions	19
Peroxisomal multifunctional protein (MFP)	Cortical	Facilitates import into peroxisomes?	28
Peroxisomal malate dehydrogenase	Cortical	Facilitates import into peroxisomes?	Freeman and Muench (unpublished results)
SPIRAL1	All arrays	MT orientation	10, 11
SPIRAL2/TOR1	All arrays	MT orientation	12, 47
TAN1	Mitotic arrays	Orientation of cell division	46
Kinesin-like calmodulin-binding protein	All arrays	Motor function	67

has been used as a starting point for studies aimed at characterizing the function of several evolutionarily conserved plant MT-binding proteins. For example, orthologs of the animal MT end-binding protein (EB1) were identified by searches of the *Arabidopsis* sequence database [36, 37]. Three AtEB1 orthologs exist and were shown to have 35–38% amino acid identity with human EB1. These studies have identified the localization of EB1 isoforms in plant cells and have provided insight into their function. In addition, EB1 has been used to mark the ends of growing MTs in studies aimed at understanding the dynamics of plant MT growth [38].

The *Arabidopsis* genome sequence database has provided important information about the MAP65 family of plant MT bundling proteins. As discussed below, MAP65 was first identified using a MT co-assembly approach [18]. MAP65 cDNAs from tobacco were then cloned and sequenced [39]. This sequence information was then used in database searches to identify nine members of the *Arabidopsis* MAP65 family [reviewed in reference 40]. These family members have molecular masses that range from 54 to 80 kDa, and they demonstrate temporal and spatial differences in their interactions with the four plant MT arrays. For instance, one family member (AtMAP65-4) preferentially localizes to the spindle poles during mitosis, whereas AtMAP65-6 co-localizes with mitochondria. This range of activities reveals the complex roles that MAPs play in various stages of the cell cycle.

Genome analysis has revealed interesting facts about other MT-binding proteins. For instance, the *Arabidopsis* genome contains 61 kinesin-related proteins [41]. In contrast, there is no clear evidence for homologs of dynein in *Arabidopsis* [42], although rice appears to have several dynein heavy-chain sequences [43].

Genetic Screens Have Identified Homologous and Plant-Specific MT-Binding Proteins

An approach aimed at identifying mutants with impaired MT organization was conducted using immunofluorescence microscopy [15]. This screen identified a temperature-sensitive *Arabidopsis* mutant that had shortened and disorganized MTs at restrictive temperatures. This mutant, *microtubule organization 1* (*mor1*), encodes an ortholog of the highly conserved MAP215/Dis1 family of MAPs. At restrictive temperatures, this mutant demonstrated severe phenotypes throughout development, including helical twisting of organs and isotropic cell expansion.

Helical twisting and isotropic cell growth phenotypes, such as those observed in the *mor1* mutant, are common in mutants that affect MT organization. The *Arabidopsis* mutant, *spirall* (*spr1*), showed a distinctive twisted growth phenotype in various organs of the plant [44]. The *spr1* locus codes for a novel 12-kDa MT-binding protein that labels all four MT arrays and is concentrated at the growing plus ends of cortical MTs [10, 11]. This protein is thought to serve as a stabilizing factor for cortical MTs, possibly by crosslinking proteins at the plus end of the MT.

Another mutant with impaired MT function showed visible defects in cell shape. Leaves of the maize mutant, *tan1*, possessed cells that divided in abnormal orientations without altering leaf shape [45]. TAN1 is another novel plant MT-binding protein that binds to MTs *in vivo*, and it is thought to have a role in guiding expanding phragmoplasts to pre-established division sites on the cell cortex [46]. Several other interesting

mutants were shown to have altered MT-binding protein activity. For example, *fragile fibre2*, a mutant that affects cell wall biosynthesis, encoded a katanin-like MT-severing protein [13], and the *spr2/tortifolia1* mutant demonstrates helical growth and codes for a novel MT-binding protein that regulates cortical MT orientation [12, 47]. The *Zwichel* gene codes for a kinesin-like CaM-binding protein (KCBP), and plants having mutations in this gene have unusual trichome phenotypes [48].

Biochemical Approaches: Fishing for MT-Binding Proteins

Numerous MT-binding proteins, some of which are plant-specific, have been identified using biochemical purification approaches. Co-assembly and co-purification of plant MT-binding proteins with MTs polymerized *in vitro* was first used to characterize the bundling activities of these proteins and to resolve their profiles by gel electrophoresis. For instance, carrot MT-binding proteins that were co-purified with taxol-stabilized neuronal MTs demonstrated MT-bundling activity [23]. Experiments followed that utilized the co-assembly of maize MT-binding and bundling proteins with endogenous maize MTs [20, 21]. The use of MAC and TAC techniques to isolate plant MT-binding proteins was first reported in the mid-1990s [27] and has since been successfully employed in other studies. These experiments laid the groundwork for subsequent studies that resulted in the purification and identification of plant MT-binding proteins.

The MAP65 family of proteins represents a class of MAP that was originally isolated through rounds of polymerization and depolymerization of endogenous plant tubulin [18]. As mentioned earlier, these proteins likely regulate a range of MT activities throughout the cell cycle [39, 40]. They are reported to possess MT-bundling activity, and they form 25- to 30-nm cross-bridges between MTs [18]. MAP65 and its mammalian and fungal orthologs have important roles in cell division, particularly in maintaining the integrity of spindle midzone MTs [reviewed in reference 40]. A similar biochemical approach using endogenous MT co-assembly purification identified a 190-kDa plant MT-binding protein that also possessed actin filament binding activity [19]. MAP190 localizes to spindles and the phragmoplast upon nuclear breakdown during mitosis, indicating that this protein may cross-link actin and MTs during cell division. In another study, a 120-kDa protein was purified from carrot protoplasts by co-sedimentation with neuronal MTs [24]. The isolated SDS-PAGE protein band was in fact composed of two related proteins that were later identified as novel plus-end-directed kinesins thought to be involved in phragmoplast activity, among other possible roles [49, 50].

TAC was used to identify three different classes of MT-binding proteins. EF-1 α was a component of a protein fraction purified using TAC of carrot cell extracts [27]. It was selectively purified from this fraction by taking advantage of its calcium-dependent binding to CaM. EF-1 α was shown by immunofluorescence labeling to bind to all four cortical MT arrays [51]. Subsequent to this, a GFP-EF-1 α fusion protein was shown to localize to cortical MTs only after the cytoplasm was artificially acidified [52]. The physiological significance of this labeling of MTs by EF-1 α under acidic conditions is unclear. Marc et al. [32] used TAC to isolate a 90-kDa PM protein that possessed MT binding activity. This protein was shown to associate with cortical MTs, and was

identified as phospholipase D (PLD) [53]. The membrane and MT-binding activities suggested that PLD functions to anchor cortical MTs to the PM, perhaps serving as a structural or signaling link between MTs and the PM. This was supported by more recent work that showed cortical MTs were released from the PM after exposure to 1-butanol, an inhibitor of phosphatidic acid synthesis by PLD [54]. Finally, TAC was used in a biochemical screen for proteins that possess MT and RNA binding activity, and it resulted in the identification of a peroxisomal matrix protein called the multifunctional protein (MFP) [28]. In addition to its MT- and RNA-binding activities, MFP possesses up to four enzymatic activities involved in the β -oxidation of fatty acids in the peroxisome. MFP was shown to bind specifically to cortical MTs in plant cells [55]. The possible significance of the MT- and RNA-binding activities of MFP in peroxisomal protein sorting are discussed below.

Large-Scale MT Proteome Studies in Plants: Two Biochemical Approaches Identify Numerous MT-Binding Proteins

Chuong et al. [4] used TAC, NEPHGE, and LC-MS/MS analysis as a large-scale strategy to identify plant MT-binding proteins. The choice of TAC as the biochemical approach to enrich for MT-binding proteins was based on previous successes using this technique [27, 28, 32], as well as on the previous observation that the SDS-PAGE profiles of protein fractions eluted from MT and tubulin columns were qualitatively similar (Figure 19.1A) [27, 28]. This study identified 122 putative MT-binding proteins, and these were classified in six categories based on predicted or known function (Figure 19.1C). MT co-sedimentation analysis demonstrated that most, if not all, of the proteins bound to MTs *in vitro*. Greater than 40% of these proteins are related to proteins that were previously shown as having an association with MTs [4].

Notable MAP-related proteins that were identified in this study were HSP70/HSP40 and a protein similar to CLASP1. Numerous RNA-binding proteins were also identified, including RNA helicases, small nuclear ribonucleoproteins, and other proteins containing known RNA-binding domains. Proteins with roles in translation were also identified, including elongation and initiation factors as well as aminoacyl-tRNA synthetases. This approach also identified several proteins with putative signaling functions, including protein phosphatase 2A (PP2A) regulatory and catalytic domain-containing proteins, a CDK, and Ser/Thr kinase domain-containing proteins. These, along with other signaling-related proteins that were identified in this study, support previous research that suggests the cytoskeleton serves as a matrix for the binding of signaling proteins [7].

This proteomic study also provided clues regarding the activation of metabolic enzymes in association with MTs in plant cells. Numerous metabolic enzymes were identified, including the glycolytic enzyme GAPDH, a protein shown previously to possess MT-binding activity [56]. Several enzymes associated with other metabolic pathways were also identified in this study. Five enzymes involved in catalyzing reactions in one-carbon metabolism were identified, as were six peroxisomal matrix proteins. Another class of proteins that were identified in the tubulin-binding protein fraction was a group of enzymes involved in cell-wall modification. This suggests that

there may be an interaction by one or more of these proteins across the PM, thereby linking the cell wall with MTs.

In another large-scale study aimed at identifying MT-binding proteins, Korolev et al. [22] identified 55 proteins that co-assembled with endogenous taxol-stabilized MTs from *Arabidopsis* suspension cell protoplasts. The efficacy of this approach was confirmed by the identification of several previously characterized MAPs and motor proteins, such as MAP65, MOR1, and several kinesins. Other nonclassical MAPs were identified, including an RNA helicase, TFs and ribosomal proteins, and the metabolic enzymes sucrose synthase and GAPDH. Several HSP subunits were also identified, including HSP70. A previously uncharacterized 70-kDa protein that was also identified was studied in more detail. This protein labeled all four MT arrays when expressed as a GFP fusion. Regions of this protein showed similarity to the rho-associated coiled-coil domain and a cdc42-binding protein kinase; however, the precise function of this plant-specific protein is not known at this time. Another protein that was identified using this same procedure was later identified as a plant-specific, 187-kDa MT-binding protein called AIR9 [57]. This protein co-localizes with cortical MTs and the pre-prophase band, and it may be involved in mechanisms that ensure that the developing cell plate contacts the pre-programmed cortical division site.

These two large-scale approaches [4, 22] were complementary in their attempts to identify MT-binding proteins. On one hand, there were several proteins that were identified in both studies, including RNA helicases, translation initiation and EFs, ribosomal proteins, HSP70, and GAPDH. However, there were many proteins that were unique to each study. For example, the MT co-assembly approach [22] identified several MAPs that are involved in MT organization, bundling, and dynamics. These proteins were not identified in the TAC approach [4], likely because these proteins bind specifically to MTs. Conversely, the TAC approach [4] identified several proteins that did not show up using the MT co-assembly approach, but have been shown to bind to MTs previously. For instance, peroxisomal matrix proteins were abundant in the tubulin-binding protein fraction, consistent with the observation that peroxisomal matrix proteins are known to interact with MTs *in vivo* [55; Freeman and Muench, unpublished results]. Additionally, several proteins with putative roles in signaling bound to the tubulin matrix, including proteins with probable phosphatase and kinase activity—proteins that have previously been shown to bind to MTs [7].

In situ localization experiments will be important in determining if the proteins identified in these studies do indeed interact with MTs in plant cells. This is a challenging task, since many of these proteins may have a low affinity for MTs or may interact transiently with MTs in the cell and would produce a low signal in subcellular localization experiments. It is also expected that a portion of the proteins that were identified in these two studies bind to MTs artifactually, and they do not represent authentic MT-binding proteins.

Possible Roles for Nonclassical MAP Interactions with MTs

The numerous nonclassical MAPs that have been identified in studies discussed here raises questions about the significance of their interactions with MTs in the cell. The

components of the plant cytoskeleton—actin filaments and MTs—appear to have defined roles in many processes in plant cells. For example, actin filaments are generally thought to serve as highways by which organelles and macromolecules move throughout the cell. Conversely, MTs likely participate in numerous anchoring events. For example, cortical MTs are known to respond to environmental stimuli, such as aluminum or low-temperature stress. This results in rapid and transient MT disassembly [reviewed in reference 3]. This disassembly of MTs may serve to release into the cytoplasm MT-bound proteins that are involved in stress signaling events. Indeed, several putative signaling proteins, such as PP2A, have been shown to interact with the plant cytoskeleton [4, 14]. The association of metabolic enzymes with MTs may play an important role in metabolic channeling. The observation that enzymes involved in one-carbon metabolism were present in the tubulin-binding protein fraction supports previous evidence for an association of these enzymes to enhance metabolic flux through the pathway [58, 59]. In addition, GAPDH, a glycolytic enzyme, has been identified as a MT-binding protein in a number of plant and non-plant studies. The interaction of this enzyme, as well as other glycolytic enzymes, with MTs may also function to assemble a metabolic channeling complex that increases the efficiency of flux through the glycolytic pathway. This type of protein concentrating mechanism provided through an interaction with MTs may also be important in translation. Numerous RNA-binding proteins and translational machinery components have been identified in MT-binding protein fractions. The cytoskeleton has long been implicated in translational regulation, and the association of proteins involved in translation may be facilitated by an interaction with MTs [60].

Dynamics of the MT Proteome: Temporal and Spatial Interactions of Proteins with MTs

The ever-expanding number of *bona fide* plant MT-binding proteins indicates that, at any one time, the MT surface is decorated with an array of proteins (Figure 19.2). The MT-binding activity of these proteins is regulated such that each may bind to one or more MT arrays, or a component of an array, as observed for members of the MAP65 proteins. Tissue- and cell-specific expression, protein phosphorylation, and competition for binding sites are known to regulate the dynamic interactions between several MAPs and MTs [61]. In plant cells, hyperphosphorylation of an isoform of MAP65 regulates its interaction with the various mitotic MT arrays in plant cells [62]. But are there other mechanisms that regulate the binding of specific proteins to MTs?

We have proposed that another possible mechanism for regulating the binding of proteins may involve translational autoregulation. Several MT-binding proteins have been shown also to possess RNA-binding activity. For example, two of the plant peroxisomal matrix proteins that were identified by TAC [4, 28] also possess RNA-binding activity [28] (Freeman and Muench, unpublished results). Additionally, mammalian GAPDH possesses RNA-binding activity in addition to its MT-binding activity [56]. We presented a model that links the MT- and RNA-binding activities of peroxisomal matrix proteins to translational autoregulation and their import into peroxisomes [63]. Peroxisomal matrix proteins are commonly imported in a fully

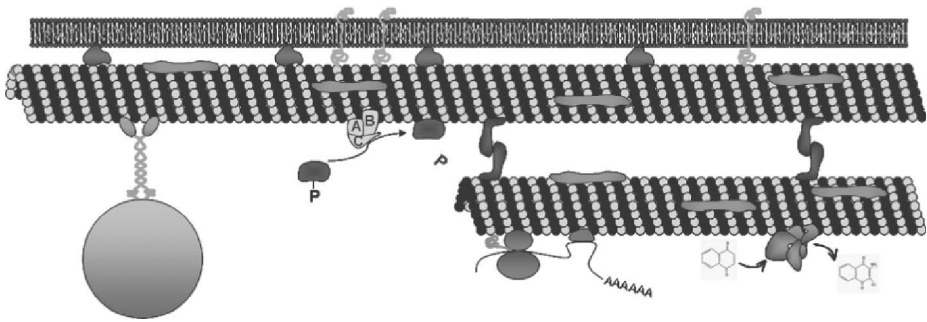


FIGURE 19.2. A model depicting possible interactions between MTs and MT-binding proteins. Structural MAPs are important in stabilizing MTs (light green), mediating the bundling of MTs (dark green), and anchoring MTs to the PM (brown). Motor proteins (turquoise) transport their cargo along MTs. Communication between MTs and the cell wall across the PM may involve TM proteins (orange). Metabolic complexes (red) may assemble on MTs to facilitate the flux of material through a metabolic pathway. Evidence for mRNA-binding protein and ribosome (violet) interactions with MTs may be important in mRNA localization and anchoring, as well as translational control. Signaling events involving protein phosphatases and kinases (yellow) could modify downstream signaling components or regulate the binding of proteins to MTs (blue). Other MT-binding proteins, such as EB1, are well-characterized in plants but are not depicted in this figure. See insert for color representation of this figure.

folded and oligomerized state and, therefore, have the potential to possess cytosolic activities prior to import. This model predicts that these enzymes interact with their cognate mRNAs while bound to MTs in order to control translation in an autoregulatory fashion. However, after the peroxisome interacts with MTs and the MT-bound peroxisomal protein is imported, translational repression is lifted. This translational autoregulation would serve as a mechanism to control the synthesis and binding of peroxisomal proteins to MTs. This model could apply to other MT-binding proteins. For example, in our tubulin-binding proteomics study [4], several plant enzymes involved in folate-dependent reactions were shown to bind to tubulin. Similar enzymes in human cells possess RNA-binding activity and utilize this RNA-binding activity to autoregulate their translation [64, 65].

19.5 CONCLUSIONS

The complex functions that are orchestrated by plant MTs necessitate that the organization of the MT network be tightly controlled. The identification of structural MAPs has progressed by leaps and bounds over the past several years. The characterization of events such as MAP phosphorylation is essential to understand the temporal and spatial control of MT organization throughout the cell cycle. As the number of structural plant MAPs continues to grow, our understanding of how MT dynamics and organization is regulated will continue to expand. It is not entirely surprising that many of these MAPs are unique to plants, since the dynamics and organization of

plant MTs differ in many ways from those of other eukaryotes. Also, there is continued support for a MT role in anchoring and concentrating metabolic enzymes, signaling proteins, and mRNA and translation complexes, presumably to regulate the activity of these proteins. The ever-increasing number of authentic MT-binding proteins has uncovered potential roles for MTs in numerous cellular processes in plants.

19.6 FIVE-YEAR VIEWPOINT

For each of the authentic MT-binding proteins identified thus far, further research will be required to better understand their role in MT organization, dynamics, and function. How do these proteins bind to MTs, and what regulates these interactions? How does MAP binding control MT organization and dynamics? Also, what role do interactions between nonclassical MAPs and MTs serve? *In situ* evidence is required to demonstrate that these interactions are in fact authentic. Genetic and biochemical tools will continue to be valuable in identifying new protein–MT interactions, as will technological advances in this field. The sensitivity and adaptability of MS techniques continues to improve. For example, MS analysis of complex protein mixtures in solution can provide abundant sequence identity information that can be used as a starting point for further MT-binding protein analysis. There is also tremendous potential for identifying new MT-binding proteins using high-throughput screening of random fusions to fluorescent proteins expressed in plant and non-plant cells. Indeed, the recent identification of plant actin-binding proteins has been achieved using a heterologous expression system [66].

ACKNOWLEDGMENTS

Our research on the plant cytoskeleton is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) awarded to DGM. We apologize to those whose work we were unable to discuss due to space limitations.

REFERENCES

1. Goddard RH, et al. 1994. *Plant Physiol* **104**:1–6.
2. Paradez A, Wright A, and Ehrhardt DW. 2006. *Curr Opin Plant Biol* **9**:571–578.
3. Wasteneys GO. 2004. *Curr Opin Plant Biol* **7**:651–660.
4. Chuong SD, et al. 2004. *Mol Cell Proteomics* **3**:970–983.
5. Mandelkow E and Mandelkow EM. 1995. *Curr Opin Cell Biol* **7**:72–81.
6. Lloyd C and Hussey P. 2001. *Nat Rev Mol Cell Biol* **2**:40–47.
7. Gundersen GG and Cook TA. 1999. *Curr Opin Cell Biol* **11**:81–94.
8. Gardiner J and Marc J. 2003. *Protoplasma* **222**:61–74.

9. Meagher RB and Fechheimer M. 2003. The Arabidopsis cytoskeleton genome, In: *The Arabidopsis Book* (Somerville CR and Meyerowitz EM, eds.), American Society of Plant Biologists, Rockville, MD.
10. Sedbrook JC, et al. 2004. *Plant Cell* **16**:1506–1520.
11. Nakajima K, et al. 2004. *Plant Cell* **16**:1178–1190.
12. Shoji T, et al. 2004. *Plant Physiol* **136**:3933–3944.
13. Burk DH, et al. 2001. *Plant Cell* **13**:807–827.
14. Camilleri C, et al. 2002. *Plant Cell* **14**:833–845.
15. Whittington AT, et al. 2001. *Nature* **411**:610–613.
16. Vallee RB. 1982. *J Cell Biol* **92**:435–442.
17. Williams R and Lee J. 1982. *Methods Enzymol* **85**:376–385.
18. Jiang C-J and Sonobe S. 1993. *J Cell Sci* **105**:891–901.
19. Igarashi H, et al. 2000. *Plant Cell Physiol* **41**:920–931.
20. Schellenbaum P, et al. 1993. *Plant J* **3**:253–260.
21. Vantard M, et al. 1991. *Biochemistry* **30**:9334–9340.
22. Korolev AV, et al. 2005. *Plant J* **42**:547–555.
23. Cyr RJ and Palevitz BA. 1989. *Planta* **177**:245–260.
24. Chan J, Rutten T, and Lloyd CW. 1996. *Plant J* **10**:251–259.
25. Kellogg DR, Field CM, and Alberts BM. 1989. *J Cell Biol* **109**:2977–2991.
26. Miller KG, et al. 1991. *Methods Enzymol* **196**:303–319.
27. Durso NA and Cyr RJ. 1994. *Plant Cell* **6**:893–905.
28. Chuong SDX, Mullen R, and Muench DG. 2002. *J Biol Chem* **277**:2419–2429.
29. O'Farrell PH. 1975. *J Biol Chem* **25**:4007–4021.
30. Nogales E. 2000. *Annu Rev Biochem* **69**:277–302.
31. Wu Y, et al. 1998. *Plant Cell Physiol* **39**:1251–1257.
32. Marc J, et al. 1996. *Plant Cell* **8**:2127–2138.
33. Balaban N and Goldman R. 1992. *Cell Motil Cytoskeleton* **21**:138–146.
34. Rutten T, Chan J, and Lloyd CW. 1997. *Proc Natl Acad Sci USA* **94**:4469–4474.
35. Sedbrook JC. 2004. *Curr Opin Plant Biol* **7**:632–640.
36. Chan J, et al. 2003. *Nat Cell Biol* **5**:967–971.
37. Mathur J, et al. 2003. *Curr Biol* **13**:1991–1997.
38. Chan J, et al. 2005. *Plant Cell* **17**:1737–1748.
39. Smertenko A, et al. 2000. *Nature Cell Biol* **2**:750–753.
40. Sasabe M and Machida Y. 2006. *Curr Opin Plant Biol* **9**:563–570.
41. Reddy AS and Day IS. 2001. *BMC Genomics* **2**:2.
42. Lawrence CJ, et al. 2001. *Traffic* **2**:362–363.
43. King SM. 2002. *Traffic* **3**:930–931.
44. Furutani I, et al. 2000. *Development* **127**:4443–4453.
45. Smith LG, Hake S, and Sylvester AW. 1996. *Development* **122**:481–489.
46. Smith LG, et al. 2001. *J Cell Biol* **152**:231–236.
47. Buschmann H, et al. 2004. *Curr Biol* **14**:1515–1521.
48. Oppenheimer DG, et al. 1997. *Proc Natl Acad Sci USA* **94**:6261–6266.

49. Barroso C, et al. 2000. *Plant J* **24**:859–868.
50. Asada T, Kuriyama R, and Shibaoka H. 1997. *J Cell Sci* **110**:179–189.
51. Durso NA, Leslie JD, and Cyr RJ. 1996. *Protoplasma* **190**:141–150.
52. Moore RC and Cyr RJ. 2000. *Cell Motil Cytoskeleton* **45**:279–292.
53. Gardiner JC, et al. 2001. *Plant Cell* **13**:2143–2158.
54. Gardiner J, et al. 2003. *Plant Cell Physiol* **44**:687–696.
55. Chuong SD, Mullen RT, and Muench DG. 2005. *BMC Cell Biol* **6**:40.
56. Sirover MA. 1999. *Biochim Biophys Acta* **1432**:159–184.
57. Buschmann H, et al. 2006. *Curr Biol* **16**:1938–1943.
58. Schirch V and Strong WB. 1989. *Arch Biochem Biophys* **269**:371–380.
59. Strong WB and Schirch V. 1989. *Biochemisrty* **28**:9430–9439.
60. Jansen R-P. 1999. *FASEB J* **13**:455–466.
61. Cassimeris L and Spittle C. 2001. *Int Rev Cytol* **210**:163–226.
62. Smertenko AP, et al. 2006. *J Cell Sci* **119**:3227–3237.
63. Muench DG and Park NI. 2006. *Can J Bot* **84**:572–580.
64. Liu X, et al. 2000. *Biochemistry* **39**:11523–11531.
65. Tai N, et al. 2004. *Front Biosci* **9**:2521–2526.
66. Abu-Abied M, et al. 2006. *Plant J* **48**:367–379.
67. Reddy AS, et al. 1996. *Plant J* **10**:9–21.

PART IV

ORGANELLE PROTEOMICS

CELL WALL

Elisabeth Jamet, Hervé Canut, Cécile Albenne, Georges Boudart,
and Rafael Pont-Lezica

20.1 INTRODUCTION

Plant cell walls are dynamic structures essential not only for cell division, enlargement, and differentiation, but also for response to environmental constraints [1–3]. They are also sources of signals for cell recognition within the same or between different organisms [4–5]. Each cell type is surrounded by a specific cell wall, leading to a great diversity of cell wall structures and compositions [1, 6]. Cell walls are natural composite structures, mostly made up of high-molecular-mass polysaccharides, proteins, and lignins, with the latter found only in specific cell types. Polysaccharides represent up to 95% of cell wall mass, whereas cell wall proteins (CWP) only account for 5–10% of that mass. Models of cell wall structure describe the arrangement of their components into two structurally independent but interacting networks, embedded in a pectin matrix [7, 8]. Cellulose microfibrils and hemicelluloses constitute the first network; the second one is formed by structural proteins such as extensins (Figure 20.1).

CWP contribute to all cell-wall functions and are essential actors in plants. Several studies exemplify unexpected roles for CWP during development or during interactions

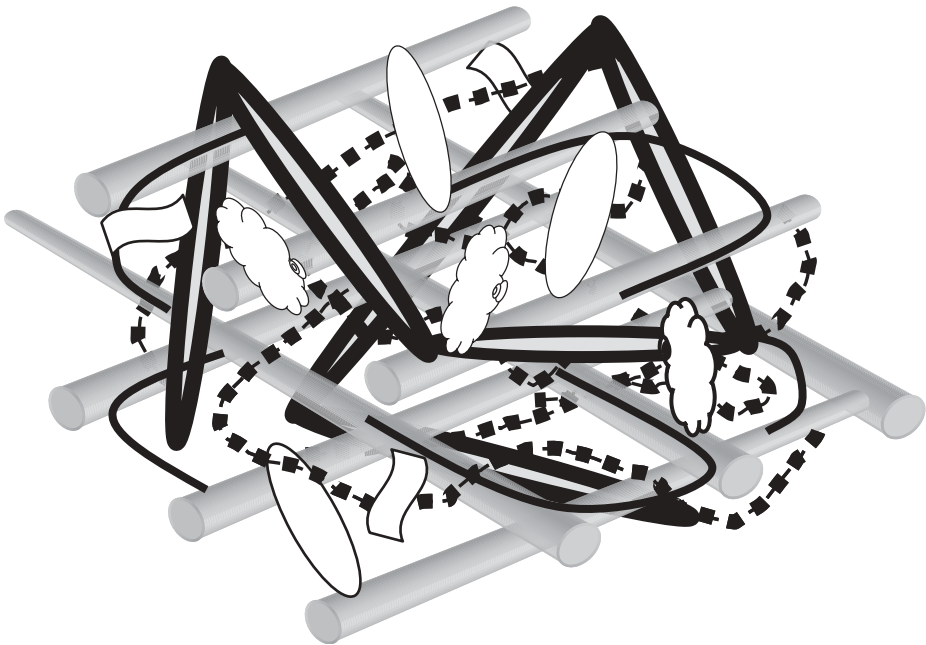









FIGURE 20.1. A model of cell wall structure including CWP. Cellulose () and hemicellulose () network interacts with structural proteins (). These two networks are embedded in a pectin matrix (). Additional proteins, classified in three types—labile proteins (), proteins weakly bound (), and proteins strongly bound () to the matrix—are also present in plant cell wall.

with pathogens. For instance, the roles of extensins have been restricted to that of structural proteins for many years. However, the phenotype of the *A. thaliana rsh* (*At1g21310*, *root-shoot-hypocotyl-defective*) mutant showed that a particular extensin is essential for correct positioning of the cell plate during cytokinesis in embryo cells [9]. Due to incorrect position of planes of division, *rsh* embryos have irregular cell shape and size, giving rise to seedlings incapable of normal development. The *A. thaliana* COBRA (*At5g60920*) CWP was shown to be essential for cellulose microfibril orientation: The *cob* mutant has severe defects in anisotropic expansion associated with disorganization of cellulose microfibrils and reduction of crystalline cellulose [10]. *PMR6* (*At3g54920*, *powdery mildew-resistant*) encodes a protein showing homology to pectate lyases and is assumed to play a key role during the infection of *A. thaliana* by *E. cichoracearum* [3]. Indeed, the *pmr6* mutant becomes resistant to infection without induction of well-known defense mechanisms. The cell wall was found to be enriched in pectins with a low degree of esterification and modified at the level of cellulose microfibril H-bonding environment. This new type of disease

resistance seems to be based on the loss of a gene required during a compatible interaction.

Extraction and analysis of CWP present specific constraints in addition to the difficulties usually encountered in proteome analysis, such as protein separation and detection of scarce proteins. Protocols should be adapted to take into account the following specificities: (i) The lack of a delimiting membrane may result in the loss of CWP during the isolation procedure, thus, ionic strength of grinding buffer should be very low; (ii) polysaccharide networks of cellulose, hemicelluloses, and pectins form potential traps for intracellular proteins that contaminate CWP, and a significant decrease in the level of CWP contamination can be obtained by working on living cells without altering their PM or by purifying uncontaminated cell walls; (iii) CWP are embedded in a polysaccharide matrix and interact in different ways with other cell wall components, making the extraction of some of them challenging, and the composition of solutions for extraction of CWP must be adapted to improve their elution from cell walls; (iv) separation of CWP by 2-DGE is not very efficient because most CWP are basic glycoproteins, and efficient alternatives to 2-DGE depend on a pre-fractionation of CWP by ion-exchange chromatography followed by separation by 1-DGE or in replacing the electrophoresis step by a direct identification of CWP by peptide sequencing using MS coupled to ionic and/or RP-LC-MS/MS; and (v) identification of heavily *O*-glycosylated CWP is very difficult using peptide mass mapping by MS, which requires introducing a deglycosylation step prior to MS analysis.

Three types of CWP can be distinguished on the basis of their interactions with cell wall components [11]. CWP can have little or no interactions with cell-wall polysaccharides or other CWP and thus move in the extracellular space. Such proteins can be found in liquid culture media of cell suspensions and seedlings or can be extracted with low ionic strength buffers. We call this fraction “labile proteins.” Most of them have acidic pI ranging from 2 to 6. Alternatively, CWP can be weakly bound to the matrix by van der Waals interactions, hydrogen bonds, and hydrophobic or ionic interactions. Such proteins can be extracted by salts. Most of them have basic pI ranging from 8 to 11, so that they are positively charged at the acidic pH of cell walls. Even though most of the cell-wall polysaccharides are neutral, pectins contain polygalacturonic acid residues that provide negative charges for interactions with basic proteins. Such interactions would be modulated by pH, degree of pectin esterification, Ca^{2+} concentration, as well as mobility and diffusion coefficients of these macromolecules [8]. Two protein domains involved in interactions with pectins have been described. A four-Arg-residue domain of a peroxidase was shown to have a high affinity for Ca^{2+} -pectate *in vitro* [12]. A domain of two polygalacturonase-inhibiting proteins comprising four residues of Arg and Lys interacts *in vitro* with pectin [13]. Finally, CWP can be strongly bound to cell wall components so that they are resistant to salt extraction. As examples, extensins are cross-linked by covalent links [14].

This chapter will provide (i) an overview of plant cell-wall proteomic studies, (ii) several strategies to analyze cell wall proteomes, (iii) a description of experimental results, and (iv) some proposals for future research.

20.2 BRIEF BIBLIOGRAPHIC REVIEW

Despite the difficulties specific to CWP extraction and analysis, cell wall proteomics has become an active field during the last years. Main results have been obtained with the model plant *A. thaliana*. However, some studies have been performed on *M. sativa* [15] and *Z. mays* [14].

Many studies have been performed on cell suspension cultures because it is an abundant material that has been widely used for analysis of composition and structure of cell walls. CWP have been extracted in two different ways: (i) without grinding, by harvesting proteins present in culture medium [17], or by extracting proteins from the surface of living cells with salt solutions [17–19]; (ii) starting with cell-wall purification, by extracting proteins with solutions containing salts, detergents, or denaturing agents [20–21].

Cell-wall proteomic studies of various plant organs have been performed: roots [16], stems [15, 22], rosette leaves [23], etiolated hypocotyls [24], and seedlings cultured in liquid medium [25]. Such studies are complementary to those performed on cell suspension cultures confirming that cell wall structure and composition are regulated during development [1, 6]. They also allow comparisons between cell wall proteomes of different organs in relation to their functions.

Due to technical limitations, proteomic studies cannot yet give an exhaustive inventory of CWP. This is the reason why certain studies have been focused on specific subproteomes that are not well-represented using global approaches. A Yariv-binding proteome was analyzed to describe arabinogalactan proteins (AGP) that are proteoglycans comprising up to 90% of polysaccharides and specifically recognized by the Yariv reagent through an interaction of the antigen–antibody type [25]. Since CWP go through the secretory pathway, most of them are supposed to be glycosylated. Concanavalin A (ConA) was used to select *N*-glycosylated proteins [22]. A subproteome was actually obtained: It was enriched in glycoside hydrolases and in multicopper oxidases, but it was missing expansins. The GAPs are located at the cell surface and can be involved in signaling or in cell adhesion. Several families of GAP were specifically isolated from lipid rafts using Triton X-114 phase partitioning and sensitivity to phosphatidylinositol-specific PLC [26]. CWP secreted in culture medium have been searched in culture medium of seedlings [27]. Interestingly, this subproteome was missing proteins having interacting domains with proteins such as LRR or with polysaccharides such as lectins.

Finally, the characterization of the *Z. mays* xylem sap proteome should be mentioned because it can reveal proteins related to different steps of xylem differentiation including transition from primary to secondary cell wall, PCD, and long-distance signaling between roots and aerial organs. Actually, 97% of the identified proteins were predicted to be secreted [28].

20.3 SPECIFIC METHODOLOGY AND STRATEGIES

The choice of a protocol to extract CWP for proteomic analysis is dependent on the plant material and on the type of proteins to be released from cell walls. Working on

living cells is probably the best solution to avoid intracellular contamination. Three types of nondestructive methods are available (Figure 20.2A): (i) analysis of liquid medium of cell suspension cultures or seedlings [17, 27]; (ii) washings of cells cultured in liquid medium with salt solutions [17, 18]; and (iii) vacuum infiltration of organs such as leaves or roots with appropriate extraction buffers [16, 23]. Both labile and weakly bound CWP can be released. When this is not possible or when a specific subproteome has to be studied, it is necessary to use a destructive method—that is, to purify cell walls or any other fraction that can contain CWP such as xylem sap or lipid rafts (Figure 20.2B). The main problem is to prevent contamination of CWP by intracellular proteins that will stick nonspecifically to cell walls. Mostly weakly or strongly bound CWP can be extracted from purified cell walls, since labile CWP are probably lost during cell wall preparation. Extracted proteins can be submitted to affinity chromatography to study a subproteome: ConA-affinity chromatography can sort out *N*-glycosylated proteins [22]; Yariv-affinity chromatography can select AGP [25].

The composition of the extraction solution is critical and determines the type of CWP that can be released from cell walls. A solution of 0.3 M mannitol infiltrated in living tissues such as leaves can solubilize only a few acidic CWP. Such proteins are expected to have no interaction with negatively charged pectins and to be located only in intercellular spaces [23]. NaCl is usually used for extraction of proteins retained by ionic interactions in cell walls. Lithium chloride (LiCl) can extract hydroxyproline-rich glycoproteins from intact cells of *C. reinhardtii* [cited in reference 11]. Calcium chloride (CaCl₂) is probably the most efficient salt for extraction of CWP [23]. The ability of acidic and neutral carbohydrates to strongly chelate calcium might explain, through a competition mechanism, that CWP weakly bound to cell wall polysaccharides can be selectively solubilized by CaCl₂. Cyclohexanediaminetetraacetic acid (CDTA), a chelating agent, solubilizes Ca²⁺-pectate. It releases a small number of proteins having domains of interaction with polysaccharides, notably proteins showing homology to lectins [23].

Once extracted, CWP can be analyzed in different ways (Figure 20.2B). A direct analysis can be performed by trypsin digestion, followed by 2D LC–MS/MS [21]. This is a very powerful technique allowing the identification of many proteins by peptide sequencing. It does not require protein separation by electrophoresis, thus avoiding the loss of certain types of CWP. An alternative is the fractionation of CWP by CAX followed by 2-DGE or 1-DGE for acidic and basic proteins, respectively. Trypsin digestion is performed in-gel prior to peptide mass mapping using MALDI–TOF–MS or peptide sequencing by LC–MS/MS [23].

Specific difficulties are encountered for hydroxyproline-rich glycoproteins (HRGP) or proline-rich proteins (PRP) that can be heavily glycosylated CWP. Up to now, only a few of these proteins have been identified in proteomic studies. Among HRGP, extensins are structural proteins forming cross-linked networks, whereas AGP can be released in cell walls after cleavage of their GPI anchor. Such proteins are poorly separated using classical techniques of electrophoresis, their biased composition in amino acids, and their high level of glycosylation prevent appropriate trypsin digestion for peptide mass mapping. However, the presence of HRGP among CWP

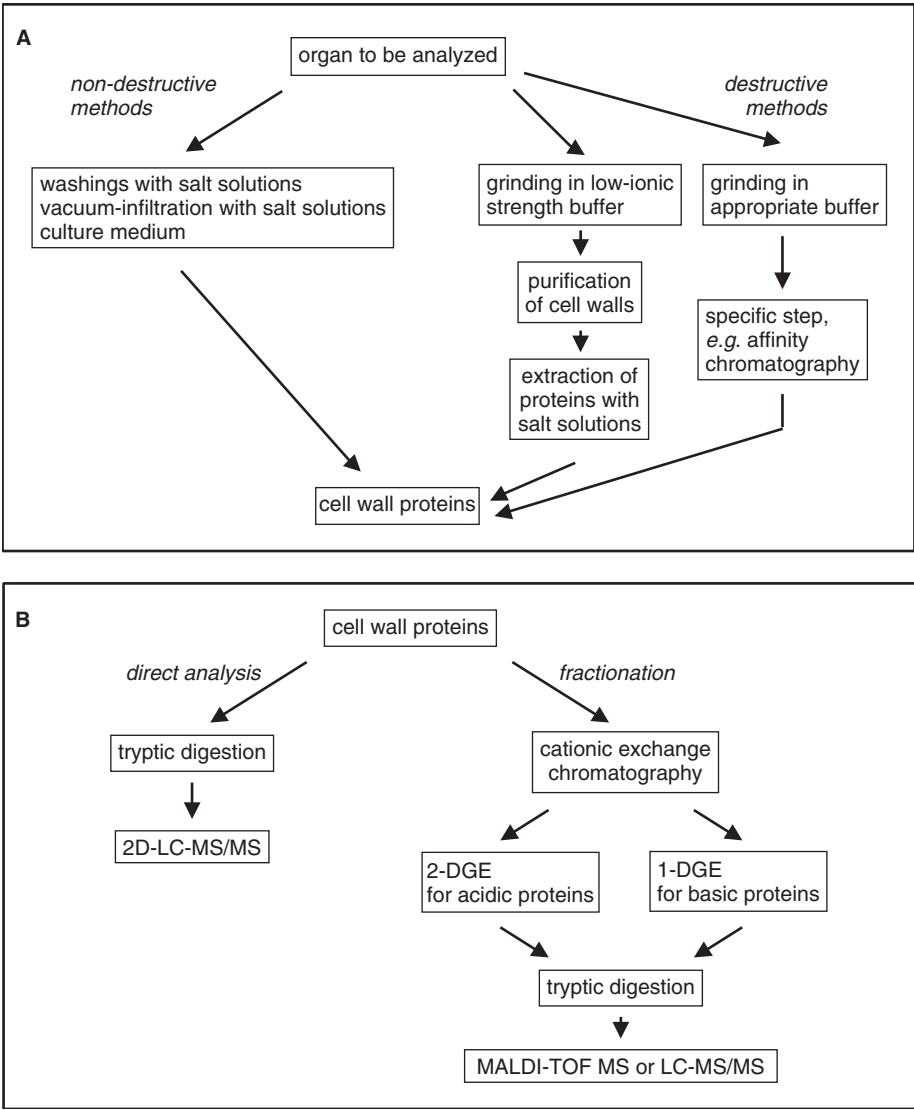


FIGURE 20.2. Strategies for extraction, fractionation and identification of CWP. **(A)** Strategies for extraction of CWP. Nondestructive or destructive methods can be chosen. A global approach can be used or a subproteome can be sorted by a specific step of purification—for example, by affinity chromatography. **(B)** Strategies for protein fractionation and identification. CWP can be directly identified using peptide sequencing by MS or fractionated by cation exchange chromatography into an acidic and a basic fraction prior to 2- and 1-DGE, respectively. Identification is performed through peptide mass mapping by MALDI-TOF-MS or peptide sequencing by LC-MS/MS.

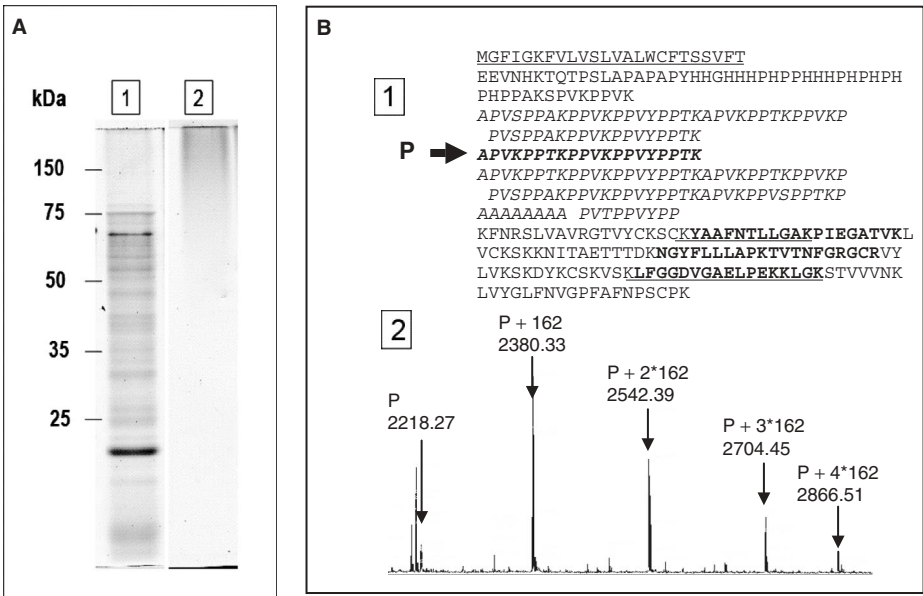


FIGURE 20.3. Analysis of cell-wall heavily glycosylated proteins. **(A)** In-gel staining of AGP with the Yariv reagent. Proteins present in culture medium of 14-day-old etiolated seedlings [27] were concentrated and separated by 1-DGE. They were stained either with colloidal CBB **(A1)** or with the Yariv reagent **(A2)**. Molecular masses of markers are indicated on the left. **(B)** Identification of a glycosylated PRP (*At1g28290*) through mass mapping using MALDI-TOF-MS. This protein was found in CWP extract from 11-day-old hypocotyls [24]. **(B1)** Amino-acid sequence of the PRP encoded by *At128290*. The predicted signal peptide is underlined. Five nonglycosylated peptides (in boldface/boldface underlined) are located in the nonrepetitive C-terminal domain of the protein allowed protein identification. A specific hydroxyproline-rich peptide (P: *APVKPPTKPPVKPPVYPTK*) from the repetitive Pro-rich domain (bold italics) was identified with five Hyp (m/z 2218.27). **(B2)** Close-up of the MALDI-TOF-MS spectrum showing modified P peptides: unmodified P (m/z 2218.27), P modified by 162.06a mass increments (m/z 2380.30 [P+162]; 2542.39 [P+2*162]; 2704.45 [P+3*162]; 2866.51 [P+4*162]) that can correspond to the addition of one to four hexose residues, respectively.

can be revealed by hydroxyproline measurement. Although AGP cannot be identified using MS without deglycosylation [26], they can be revealed after 1-DGE by staining with the Yariv reagent (Figure 20.3A). Some PRP could be identified using MALDI-TOF-MS through peptide mass mapping of their nonglycosylated parts (Figure 20.3B1). In that case, it was also possible to identify modified peptides carrying hydroxyproline residues and sugars on MALDI-TOF-MS spectra (Figure 20.3B2). A more powerful method for protein identification consists in a preliminary step of deglycosylation with anhydrous HF to remove O-linked sugars [25].

The reliability of protein profiling for a compartment like the cell wall strongly depends on the quality of the extraction protocol. The classical methods to check the

purity of a particular fraction are not conclusive for proteomic studies, since the sensitivity of the analysis by MS is 10–1000 times higher than enzymatic or immunological tests using specific markers. There have been many discussions about the possibility to find noncanonical proteins in cell walls—that is, proteins known or predicted to be intracellular [11, 29]. The most efficient way to evaluate the quality of a protocol of CWP extraction is (i) to identify all the proteins by MS and (ii) to perform extensive bioinformatic analysis to determine if the identified proteins contain a signal peptide but contain no retention signals for other cell compartments. Several programs should be used to ensure a reliable prediction: PSORT allows predicting any subcellular localization (<http://psort.ims.u-tokyo.ac.jp/form.html>); TargetP looks for the presence of signal peptides for protein secretion or transit peptides for mitochondrion or chloroplast targeting (<http://www.cbs.dtu.dk/services/TargetP/>). It is then possible to conclude about the quality of the extraction protocol by calculating the ratio of predicted secreted proteins to intracellular ones.

20.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Results of proteomics studies are usually shown as tables of raw data, making their interpretation very difficult. Prediction of protein function or functional domains by bioinformatics is a very powerful tool. It allows classifying CWP, inferring biological or biochemical functions performed by CWP, and pointing out proteins yet unknown that can perform new functions. Several programs have to be used to get reliable information: BLAST allows us to find homologous proteins and to describe multigene families (<http://www.ncbi.nlm.nih.gov/BLAST/A>); ScanProsite looks for patterns or motifs stored in the PROSITE database among which motifs with high probability of occurrence that can be useful to look for putative sites of PTM (<http://www.expasy.ch/tools/scanprosite/>); InterProScan searches for motifs or profiles stored in several databases (<http://www.ebi.ac.uk/InterProScan/>). Using such tools, an arrangement of proteins in functional classes was proposed to get an overview of the *A. thaliana* cell-wall proteomes [11]. Of course this classification has to evolve to take into account new experimental results obtained by biochemical or genetic approaches. However, the assumption is that proteins sharing conserved domains have the same activity. It should be noted that the biochemical function of only a small portion of the identified proteins was experimentally demonstrated. A CWP database was constructed [11] and now includes 365 proteins coming from several proteomics studies performed on *A. thaliana* [17–27].

About 88.5% of these CWP could be distributed in seven categories on the basis of predicted biochemical or biological functions (Figure 20.4). The remaining 11.5% consist of proteins of yet unknown function, some of them being only present in plants. Proteins acting on polysaccharides are the most abundant proteins (27.4%). They are glycoside hydrolases, glycoside transferases, carbohydrate esterases, carbohydrate lyases, and expansins. Two functional classes of CWP are of equal importance: Oxido-reductases (13.7%) include POX, multi-copper oxidases, berberine-bridge enzyme (S)-reticulon:oxygen oxido-reductases and germins; proteases

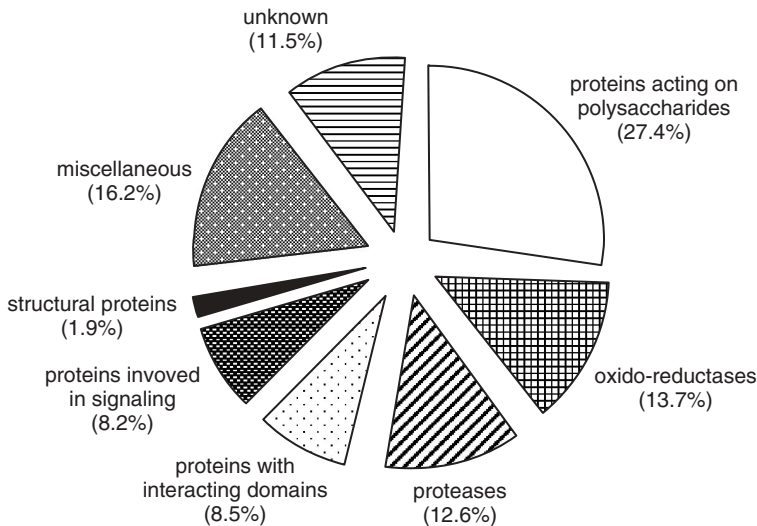


FIGURE 20.4. Distribution of *A. thaliana* CWP identified through proteomic studies in functional classes. The database comprises 365 proteins from several proteomic studies [17–27]. Protein sequences were analyzed with several bioinformatic programs to look for functional domains and homologies to proteins of known functions. They were sorted into eight functional classes: proteins acting on polysaccharides, oxido-reductases, proteases, proteins with interacting domains, proteins involved in signaling, structural proteins, proteins of unknown function, and miscellaneous. The percentage of proteins in each functional class is indicated in parentheses.

(12.6%) are of several types, namely, subtilisins, aspartate proteases, cysteine proteases, and serine carboxypeptidases. Proteins having interacting domains (8.5%) include proteins interacting with proteins through LRR domains, lectins interacting with sugars, enzyme inhibitors such as polygalacturonase inhibitor proteins (PGIP), pectin methylesterases inhibitors (PMEI), and protease inhibitors. Proteins involved in signaling (8.2%) are mainly AGP and LRR-receptor protein kinases that have been identified through their extracellular LRR domains. Other proteins of various functions (16.2%) were put together in a functional class called “miscellaneous,” among which are proteins homolog to acid phosphatases, Bcp, and proteins having lipase/acyl hydrolase domains. These proteins are awaiting additional experimental data to be more precisely classified.

Are proteomes exhaustive descriptions of the protein content of cell walls? As mentioned above, global approaches cannot lead to the extraction of all proteins from cell walls due to the diversity of protein structures, abundance, and interactions with cell wall components. Additional limitations are separation and analysis techniques. Each proteome is thus missing CWP. A first example is provided by the rosette proteome that is lacking POX, whereas POX activities have been found in rosette leaves [23]. The use of several salt solutions to release CWP by vacuum infiltration was probably not sufficient to elute POX that can be bound to Ca^{2+} -pectates [13].

As a second example, structural proteins are underrepresented in all cell wall proteomes. Either those proteins were not extracted from cell walls due to their cross-linking in networks through covalent links [12] or they were not identified using MS due to their high level of glycosylation.

Do proteomes take into account PTM? Although trains of spots separated by 2-DGE are mentioned in many studies, only a few experimental data are available to show that identified proteins can be isoforms differing by PTM. Glycosylation of several CWP was visualized by fluorescent staining with ProQ Emerald[®] 300 (Pro-Q EPS, Molecular Probes): As many as 12 and 17 isoforms were stained for a subtilisin serine protease and a PGIP, respectively [28]. The presence of *N*-glycosylations was shown on the COBRA protein by a shift in electrophoretic mobility following a treatment with a peptide *N*-glycosidase (10). The presence of a GPI anchor on COBRA was shown by treatment of membranes with a phosphatidylinositol-specific PLC that led to a release of the protein in the aqueous phase [10]. Phosphorylation of a lectin (At1g78850) and a chitinase (At3g12500) was shown after separation by 2-DGE and positive reaction to antibodies against pTyr residues [30].

Are the same proteins found in the cell walls of different plant organs? It is difficult to answer this question for several reasons. Since quantitative data are missing, comparisons should rely on qualitative data—that is, presence/absence of a protein. Proteomes should be obtained using comparable protocols for extraction, separation, and identification of proteins by MS. Differences in polysaccharide composition, cell-wall structure, a lower abundance of the protein, or PTM might be responsible for the absence of a protein in a proteome. However, taking into account all these restrictions, a comparison between several proteomes of *A. thaliana* is proposed in Table 20.1: 4- to 5-week-old rosettes [23], stems at the late flowering stage [22], cell suspension cultures [17, 19–21], and 11-day-old hypocotyls [24]. It is noteworthy that despite the presence of proteins of all functional classes in each proteome, a great part of each proteome (between 36.5% and 43.6% of identified proteins) is specific. These differences are mainly due to the presence of different members of the same protein families which might be differently regulated during development. The case of oxido-reductases is shown in Table 20.2. As many as 23 POX were identified as well as eight multicopper oxidases and six berberine-bridge (*S*)-reticulin:oxygen

TABLE 20.1. Specificities of *A. thaliana* Cell-Wall Proteomes^a

Organ (Age)	Proteins Shared by at Least Two Proteomes (%)	Specific Proteins (%)
Rosettes (4–5 weeks)	63.5	36.5
Stems (late flowering stage)	58.0	42.0
Cell suspension culture	56.4	43.6
Etiolated hypocotyls (11 days)	64.4	35.6

^aResults are expressed as percentages of total number of proteins identified in each proteome: 4- to 5-week-old rosettes [23], stems at the late flowering stage [22], cell suspension cultures [17, 19–21, 26], 11-day-old etiolated hypocotyls [24].

TABLE 20.2. Origin of the Specificity of Cell Wall Proteomes: Some Examples Taken from *A. thaliana* Multigene Families of Oxido-reductases^a

Gene	Protein	Rosettes (4–5 weeks)	Stems (Late Flowering Stage)	Cell Suspension Cultures	Etiolated Hypocotyls (11 days)	Culture Medium of Etiolated Seedlings (14 days)
Peroxidases						
At1 g05240	AtPrx01			1		
At1 g49570	AtPrx10					1
At1 g71695	AtPrx12			1		
At2 g18140	AtPrx14			1	1	
At2 g18150	AtPrx15			1		1
At2 g38380	AtPrx22		1			1
At3 g01190	AtPrx27					1
At3 g03670	AtPrx28		1			
At3 g21770	AtPrx30			1	1	
At3 g28200	AtPrx31			1		
At3 g32980	AtPrx32			1	1	
At3 g49110	AtPrx33		1			
At3 g49120	AtPrx34		1	1	1	1
At4 g08770	AtPrx37			1		
At4 g30170	AtPrx45				1	1
At4 g36430	AtPrx49			1		
At5 g05340	AtPrx52			1		
At5 g06720	AtPrx53			1		1
At5 g17820	AtPrx57			1		
At5 g42180	AtPrx64			1		
At5 g64100	AtPrx69			1		
At5 g64120	AtPrx71			1		1
At5 g66390	AtPrx72				1	
Multicopper Oxidases						
At4 g25240	SKS1			1		
At4 g22010	SKS4		1		1	
At1 g76160	SKS5	1	1	1	1	
At1 g41830	SKS6	1	1	1		
At1 g21860	SKS7		1			
At4 g38420	SKS9			1		
At5 g66920	SKS17		1			
At4 g12420	SKU5		1	1		
Homologs to Berberine-Bridge Enzyme (S)-Reticulin:Oxygen Oxido-reductase						
At1 g30730						
At2 g34790		1		1		
At4 g20830			1	1		
At5 g44380				1	1	
At5 g44390				1		
At5 g44400		1				

^aThree families of oxido-reductases are analyzed: peroxidases, multicopper oxidases, and homolog to berberine-bridge (S)-reticulic: oxygen oxido-reductases. The same proteomes as in Table 20.1 are compared.

oxido-reductases. Some of them were common to several proteomes; others were specific to a single one. Only six proteins were found in the four proteomes compared: the XYL1 alpha-xylosidase (At1g68560), the AtXTH4 xyloglucan endotransferase (At2g06850), the SKS5 multicopper oxidase homolog to SKU5 (At1g76160), two lectins with curculin-like domains (At1g78850, At1g78860), and a protein homolog to the tomato xyloglucan-specific endoglucanase inhibitor protein (XEGIP, At1g03220).

20.5 CONCLUSIONS

Proteomics greatly contributed to a better knowledge of CWP that are believed to play crucial roles in cell wall structure, cell wall polysaccharide modifications, plant defense, signaling, and lignification. The organ-specificity of cell wall proteomes can be related to the diversity of cell walls and to the complexity of their composition and structures. The great diversity of CWP, and especially of glycoside hydrolases, found in rosette and stem proteomes was not expected. It suggests a great plasticity of cell walls, even in well-differentiated tissues. Proteomics identified new proteins of yet unknown functions; some of them seem to be major cell wall proteins. The characterization of PTM in relation to protein function has already started thanks to progress in MS technologies. The presence of phosphatases, proteases, and glycoside hydrolases suggests a complex regulation of CWP involving various types of post-translational processing events such as dephosphorylation and hydrolytic processing by proteases or glycosidases. Finally, the identification of different types of proteins contributing to the same physiological process should help understanding cell wall functions.

20.6 FIVE-YEAR VIEWPOINT

Although a lot of information on cell wall proteomes is now available, many questions concerning CWP remain unanswered. They concern the exhaustive description of proteomes, the regulation of gene expression through PTM and protein degradation, the role of the cell wall in signaling via oligosaccharides or oligopeptides generated by glycoside hydrolases or proteases, the precise biological function of proteins for which a biochemical function is predicted, and the role of proteins of yet unknown function.

Saturation of the cell wall proteome is not achieved yet. From bioinformatics annotations of the *A. thaliana* genome and the work of the Cell Wall Genomic Group of Purdue University (IN, USA) (<http://cellwall.genomics.purdue.edu/families/index.html>), the number of CWP can be estimated between 1000 and 2000 proteins [11]. We already listed 365 CWP from proteomic studies, meaning that about one fourth of CWP has already been identified. Several strategies could be used to increase the saturation of the cell wall proteome. Separation of CWP should be improved prior to identification by MS. Steps of chromatography could be introduced prior to 1-DGE for a better fractionation. MS analysis could also be done on peptide mixtures directly obtained from CWP to skip the electrophoresis step that appears

to be limiting. Strongly bound proteins such as structural proteins are still lacking. Big efforts should be made to extract and to analyze those CWP that are physically linked to cell wall components. Enzymes or chemicals could be used to degrade cell wall polysaccharides while maintaining protein integrity. Either new types of CWP should be released or the same types of proteins as with salt solutions but linked in other ways to cell wall components. Finally, an *N*- or *O*-deglycosylation step could be introduced prior to protein analysis to allow identification of highly glycosylated proteins such as HRGP.

A comprehensive understanding of gene regulation requires considering all steps from gene transcription to protein degradation. More and more, comparisons of data of transcriptomics and proteomics show that the amount of an mRNA is not strictly correlated to that of the translated protein [31]. Protein degradation is rarely taken into account, although it is certainly an essential step considering the high number of proteases among CWP. Many techniques of protein quantification are available for accurate comparison of proteomes of different plant organs or from various physiological stages. Since the cell wall proteome cannot be resolved properly by 2-DGE, MS should be used after incorporation of stable isotopes either *in vivo* through metabolic labeling or post-harvest using chemical approaches [32]. None of these techniques have yet been applied to cell wall proteome probably because they are not very easy to handle. Moreover, PTM can greatly affect protein function. Since CWP are frequently glycosylated, *N*- and *O*-deglycosylation protocols should be improved. The development of MS techniques should facilitate the analysis of PTM, thus increasing our knowledge of CWP structures in relation to protein activity [33].

In addition to its great contribution to the knowledge of CWP, proteomics can contribute understanding CWP functions. Cell walls are assumed to play key roles in signaling during development and in response to environmental constraints. First, the molecular basis for such roles was provided by the discovery of oligosaccharides named elicitors. Many glycoside hydrolases are present in all proteomes (21.1% of CWP) including those from “fully differentiated” organs. Looking for substrates of such enzymes would give a lot of information on oligosaccharides mobile in cell walls. Proteases are assumed to release oligopeptides—now considered as peptide hormones, most of which are extracellular [34]. As an example, the *A. thaliana* SDD1 (*At1g04110*, *stomatal density and distribution*) cell wall protease was shown to be involved in stomatal patterning [35]. The *sdd1-1* mutant exhibits stomata clustering and increased stomatal density. It is assumed that SDD1 generates an extracellular signal regulating the number of asymmetric divisions in satellite meristemoids. The challenge is now to identify the substrates of such proteases and to characterize released protein fragments as well as their targets. Indeed, oligopeptides below 10 kDa still escape proteomic analysis.

The biological functions of the majority of the identified proteins have not yet been experimentally studied. Proteomics give information on the presence/absence of the protein in an organ or in response to environmental constraints and on PTM eventually essential for its functionality. Bioinformatic predictions provide useful clues to design relevant experiments for understanding its biochemical and biological functions. When proteins are encoded by orphan genes, no hint can be inferred.

In all cases, a full description of CWP biological functions will require complementary approaches including genetics, biochemistry, study of pattern of expression, and immunocytochemistry. Unexpected cell wall functions will probably arise from these studies.

The importance of protein/polysaccharide and PPIs has been neglected up to now to understand supramolecular assembly of cell wall components. Actually, bioinformatic analysis of weakly bound CWP identified by proteomic studies showed that about 8.5% of them have domains of interaction with proteins or polysaccharides. Moreover, the importance of cell-wall-modifying enzymes in all proteomes (27.4%) was probably not anticipated. All these proteins contribute to cell wall structure and modifications during development or in response to environmental constraints. They are candidates to design new biotechnological tools to get cell walls with modified structures for industrial purposes.

ACKNOWLEDGMENTS

The authors are grateful to the Université Paul Sabatier (Toulouse III, France) and the CNRS for support.

REFERENCES

1. Roberts K. 2001. *Plant Physiol* **125**:127–130.
2. Ellis C, et al. 2002. *Plant Cell* **14**:1557–1566.
3. Vogel JP, et al. 2004. *Plant J* **40**:968–978.
4. Pennell R. 1998. *Curr Opin Plant Biol* **1**:504–510.
5. Brownlee C. 2002. *Curr Opin Plant Biol* **5**:396–401.
6. Freshour G, et al. 1996. *Plant Physiol* **110**:1413–1429.
7. Carpita N and Gibeaut D. 1993. *Plant J* **3**:1–30.
8. Cosgrove DJ. 2005. *Nat Rev Mol Cell Biol* **6**:850–861.
9. Hall Q and Cannon MC. 2002. *Plant Cell* **14**:1161–1172.
10. Roudier F, et al. 2005. *Plant Cell* **17**:1749–1763.
11. Jamet E, et al. 2006. *Trends Plant Sci* **11**:33–39.
12. Carpin S, et al. 2001. *Plant Cell* **13**:511–520.
13. Spadoni S, et al. 2006. *Plant Physiol* **141**:557–564.
14. Brady JD, Sadler IH, and Fry SC. 1996. *J Biochem* **315**:323–327.
15. Watson BS, et al. 2004. *Phytochemistry* **65**:1709–1720.
16. Zhu J, et al. 2006. *Plant Physiol* **140**:311–325.
17. Borderies G, et al. 2003. *Electrophoresis* **24**:3421–3432.
18. Kwon HK, Yokoyama R, and Nishitani K. 2005. *Plant Cell Physiol* **46**:843–857.
19. Robertson D, et al. 1997. *J Biol Chem* **272**:15841–15848.
20. Chivasa S, et al. 2002. *Electrophoresis* **23**:1754–1765.
21. Bayer EM, et al. 2006. *Proteomics* **6**:301–311.

22. Minic, Z. 2006. Laboratoire de Biochimie des Signaux Régulateurs Cellulaires et Moléculaires, FRE2621 CNRS, Paris. Personal communication.
23. Boudart G, et al. 2005. *Proteomics* **5**:212–221.
24. Feiz L, et al. 2006. *Plant Methods* **2**:10.
25. Schultz CJ, et al. 2004. *J Biol Chem* **279**:45503–44511.
26. Borner GH, et al. 2003. *Plant Physiol* **132**:568–577.
27. Charmont S, et al. 2005. *Phytochem* **66**:453–461.
28. Alvarez S, et al. 2005. *J Proteome Res* **5**:963–972.
29. Slabas AR, et al. 2004. *Biochem Soc Trans* **32**:524–528.
30. Ndimba BK, et al. 2003. *Proteomics* **3**:1047–1059.
31. Kolkman A, et al. 2006. *Mol Syst Biol* **2**:2006–2026.
32. Ong S-E and Mann M. 2005. *Nature Chem Biol* **5**:252–262.
33. Morelle W, et al. 2006. *Proteomics* **6**:3993–4015.
34. Matsubayashi Y and Sakagami Y. 2006. *Annu Rev Plant Biol* **57**:649–674.
35. von Groll U, Berger D, and Altmann T. 2002. *Plant Cell* **14**:1527–1539.

PLASMA MEMBRANE: A PECULIAR STATUS AMONG THE CELL MEMBRANE SYSTEMS

Geneviève Ephritikhine, Anne Marmagne, Thierry Meinel, and
Myriam Ferro

21.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Biological Overview

The Plasma membrane (PM) in plant cells plays a specific role because it represents the first barrier between the surrounding environment and the interior of the cell. Therefore, PM displays specific features compared to all the other membrane systems. Basically, PM can be defined by four main characteristics. First of all, its role as an exchange barrier for ions and solutes from the soil solution has to be pointed out, thus ensuring nutrition and signaling functions. The PM also participates in protein exchanges between external environment and intracellular medium, because it is the final target of secretory pathways and the first target of internalisation processes. Likewise, PM is specifically the site for signal perception and early transduction events, whatever they originate from (biotic and abiotic signals, phytohormones, peptides, etc.), during growth and development. With regard to the PM position at the interface cell wall/cytosol, PM is a site for enzyme activities generating electron fluxes and synthesizing cell wall constituents. Being the site of many specific functions, the PM

proteome is expected to gather many different types of proteins with regard to their physicochemical properties.

The Challenges for PM Proteomics Investigations

Although they represent almost one-third of eukaryotic genomes, membrane proteins are more difficult to analyze than soluble proteins and their representation in data sets is generally low. This is especially true with integral membrane proteins for the following reasons: (i) Due to physicochemical heterogeneity of these proteins (often they are basic), 2D gel separation is not appropriate for a comprehensive mapping of membrane proteins; many hydrophobic proteins are not solubilized in the nondetergent IEF sample buffer and precipitate at their pI; and (ii) low-abundance proteins, including rare membrane proteins, are out of the scope of standard proteomic techniques [reviewed in references 1 and 2]. In the context of plant membrane proteomics studies, a pioneering effort resulted in the first catalogue of proteins present in the PM isolated from leaves of *Arabidopsis* plants [3, 4]. However, while based on the extraction and solubilization of proteins using various detergents and the use of 2-DGE, the majority of proteins identified in these studies corresponded to peripheral proteins. Proteomics approaches have benefited from the major advances achieved in the *Arabidopsis* genome sequencing project, as well from methodological progress in sensitive and rapid identification of proteins by MS and in the use of various extraction and separation procedures. The present chapter will deal with those recent PM proteomics studies developed to analyze the complex functions of this membrane system in *Arabidopsis* (Table 21.1).

21.2 SPECIFIC METHODOLOGIES AND STRATEGIES (BOX 21.1)

Plant PM Purification

The Biological Material: Cell Suspensions Versus Whole Plants or Tissues. Most of the PM proteomes reported so far concern PM fractions isolated from *Arabidopsis* cell suspensions (Table 21.1). Although this material represents a limited physiological interest compared to whole plants, three main reasons may account for this choice: (i) It produces a large amount of starting material for grinding, (ii) it allows PM protein recovery with the highest yield, and (iii) it gives a highly pure PM fraction (Table 21.1 and references herein). Indeed, cell suspensions are per se a homogeneous material, poor in chloroplasts, one of the possible major contaminants. Moreover, cell plasmolysis by an osmotic shock is easy to perform prior to cell breakage and is especially efficient on cell suspensions, thus favoring subcellular fractionation for PM extraction. It is also interesting to emphasize the interest of cell suspensions as a biological tool for concentrating proteins involved in specific functions such as protein maturation, protein turnover, cell division, cell differentiation, and so on. Almost all the plant PM proteomics studies have been performed on the model plant *A. thaliana* (Table 21.1) for two main reasons: the access to its complete sequenced genome and the development of powerful bioinformatics tools for genome

TABLE 21.1. A Brief Review of the PM Proteome Literature^a

Authors	Year	Starting Material	PM Fraction Characterization	Specificity	Strategy	Electrophoresis/Peptides Retrieval	MS	N ₀ of Proteins
Alexanderesson et al. [12]	2004	<i>A. thaliana</i> leaves and petioles	n. r.	Plasma membrane	Two-phase partitioning, salt and detergent treatment	1D-SDS-PAGE	LC-MS/MS	238
Borner et al. [26]	2003	<i>A. thaliana</i> callus cultures	n. r.	Lipid rafts	Two-phase partitioning and detergent-insoluble membrane preparation	1D-SDS-PAGE and 2D-DIGE	LC-MS/MS	30
Borner et al. [23]	2005	<i>A. thaliana</i> callus suspension	n. r.	Lipid rafts	Two-phase partitioning and detergent-insoluble membrane preparation (Triton X-100)	1D-SDS-PAGE and 2D-DIGE	LC-MS/MS	45
Dunkley et al. [33]	2006	<i>A. thaliana</i> callus cultures	n. r.	Global organelle proteomics	Membrane fractionation and iTRAQ labelling (LOPIT)	In-solution digestion (TEAB buffer)	2D-LC-MS/MS	92
Elortza et al. [20]	2003	<i>A. thaliana</i> cell suspension	n. r.	GPI-anchored proteins	Two-phase partitioning and GPI-specific PLC purification	1D-SDS-PAGE	LC-MS/MS	44
Elortza et al. [27]	2006	<i>A. thaliana</i> cell suspension	n. r.	GPI-anchored proteins	Two-phase partitioning and GPI-specific PLD purification	1D-SDS-PAGE and in-solution digestion (LysC)	LC-MS/MS	35
Lanquar et al. [19]	2007	<i>A. thaliana</i> cell suspension	Enzymatic and immunological markers	Plasma membrane, quantitative study	Two-phase partitioning, solvent extraction and salt treatments	1D-SDS-PAGE	LC-MS/MS	25

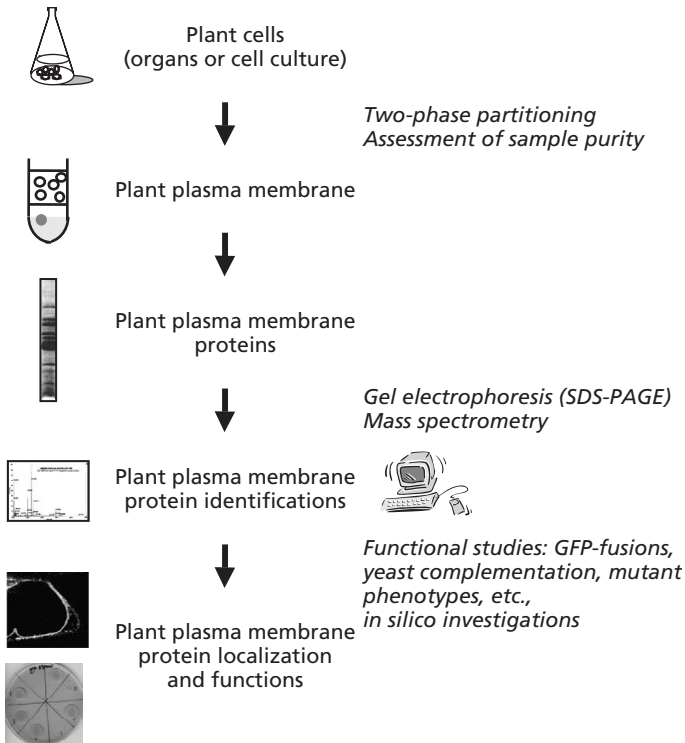
(continued)

TABLE 21.2. (Continued)

Authors	Year	Starting Material	PM Fraction Characterization	Specificity	Strategy	Electrophoresis/ Peptides Retrieval	MS	N _b of Proteins
Marmagne et al. [10]	2004	<i>A. thaliana</i> cell suspension	Enzymatic and immunological markers	Plasma membrane	Two-phase partitioning, solvent extraction and salt treatment (NaOH)	1D-SDS-PAGE	LC-MS/MS	104
Marmagne et al. [13]	2007	<i>A. thaliana</i> cell suspension	Enzymatic and immunological markers	Plasma membrane	Two-phase partitioning, salt treatments (NaCl, Na ₂ CO ₃)	1D-SDS-PAGE	Nano-LC-MS/MS	446
Mongrand et al. [5]	2004	Tobacco leaves and BY-2 cells	Enzymatic and biochemical markers	Lipid rafts	Two-phase partitioning and detergent-insoluble membrane	1D-SDS-PAGE	LC-MS/MS	~200
Morel et al. [6]	2006	Tobacco BY-2 cells	Enzymatic and biochemical markers	Lipid rafts	Two-phase partitioning and detergent-insoluble membrane	1D-SDS-PAGE	LC-MS/MS	145
Nelson et al. [14]	2006	<i>A. thaliana</i> cell suspension	Enzymatic and immunological markers	Plasma membrane	Two-phase partitioning	1D-SDS-PAGE	LC-MS/MS	70 (174)
Niuhse et al. [15]	2004	<i>A. thaliana</i> cell suspension	n. r. markers	Phosphorylated membrane proteins	Salt treatments and enrichment in phosphopeptides	Direct digestion of the membranes + IMAC enrichment	LC-MS/MS	~200
Santoni et al. [3]	1998	<i>A. thaliana</i> leaves	n. r.	Plasma membrane	Two-phase partitioning	2DE	Edman sequencing	82
Santoni et al. [4]	1999	<i>A. thaliana</i> leaves	n. r.	Plasma membrane	Two-phase partitioning	2DE	Edman sequencing	104

"For comparison, the main features characterizing every reported proteome are listed. The number of proteins retrieved is indicated in the last column; in the Nelson et al. report, [14], 70 proteins out a total of 174 were enriched in the PM fraction. n.r.: the characterization of the PM fraction was not reported by the authors. Proteins identified in Santoni et al. [3, 4] have not been included in the compilation.

Box 21.1 Outline of complementary strategies for the proteomic investigation of plant plasma membrane proteins



annotation. A few recent reports deal with tobacco plants for which the interpretation of MS data are more limited because they rely in that case on sequence homologies with *Arabidopsis* or other plant proteins [5, 6].

The Methodologies for PM Purification. A few methods are routinely used for PM purification, all starting from microsomal fractions prepared by a series of differential centrifugation applied on homogenized tissues (Figure 21.1). The PM purification procedure depends on the final objective. Biochemical studies aiming at the comparison between properties displayed by several membrane systems rely in most cases on the use of sucrose gradients; this is the most efficient technique to split microsomes in their different membrane components on the basis of their density (PM, tonoplast, Golgi membranes, etc.). At the opposite, the major challenge for proteomics studies devoted to PM is to get rid of all the other cell membrane systems (endomembranes) in order to analyze a highly pure PM fraction. The two-phase partitioning in aqueous polymers—a mixture of dextran and PEG—is the most suitable technique (Figure 21.1), because the PEG upper phase is highly enriched in PM while all the

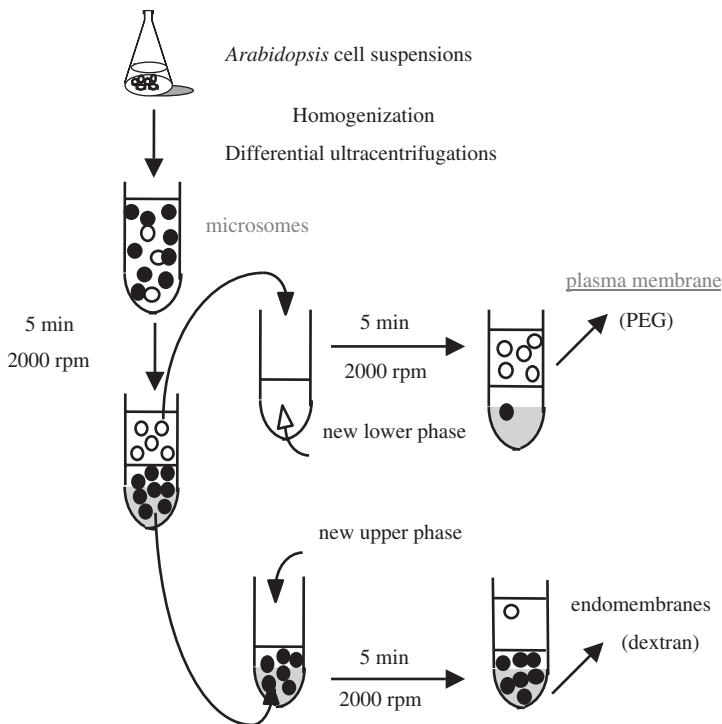


FIGURE 21.1. The two-phase partitioning method (PEG/Dextran) to isolate a PM-enriched fraction. The PM vesicles are concentrated in the PEG upper phase; all the other endomembranes are recovered in the dextran lower phase. The fraction enriched in PM will be pelleted by an ultracentrifugation (for details, see reference 8)

endomembranes are concentrated in the lower phase [7]. Almost all the PM proteomes reported so far showed the use of the two-phase partitioning method to purify PM fractions (Table 21.1). Two recent reviews discussed the advantages of that procedure and described its use for proteomics approaches [8, 9].

How to Assess Sample Purity? Historically, in order to estimate both the yield of PM protein recovery and the contamination from other subcellular compartments, enzymatic markers known from former biochemical studies as showing specific activities assigned to a membrane system were measured. Typically, P-type H^+ ATPase activity was used as a specific marker of the plant PM [10], together with markers from other plant membrane system [reviewed in reference 11]. During the past decade, the number of well-characterized membrane proteins increased significantly, providing new tools for detecting individual proteins and consequently for characterizing a membrane fraction. Enrichment of immunoreactivity for known marker proteins of the target subcellular compartments, together with a decrease in the immunoreactivity for markers known to be part of other subcellular compartments, can be monitored. To date, complementary immunological tests are currently

carried out on membrane fractions to assess both PM enrichment and PM purity (Table 21.1).

Still nowadays, endomembrane contaminants in the PM suspension cannot be avoided, because none of the available extraction/purification procedures allow the recovering of a 100% pure PM fraction so far. Among the advantages, the proteomics analyses themselves bring an overview of the contamination level, and not a view restricted to one, at most two proteins per endomembrane. Most of the reported PM proteomes point out as major contaminations (i) ribosomal proteins [12, 13] and (ii) proteins from other cell compartments, and especially the cytosol, that can be assessed by quantitative proteomics approaches [14]. Considering results combining both biochemical and enzymological tests on a same membrane fraction, contamination was assessed to be in the range of a few percent (Table 21.1 and references therein).

Specific Recovery of Plant PM Proteins: Complementary Methods

Pioneering efforts to investigate the plant PM proteome were undertaken using 2-DGE [3, 4]. However, while based on an optimized solubilization using detergents, the 2-DGE-based separation allowed mainly peripheral proteins to be identified. In the following section, we will describe recent developments, using SDS-PAGE separation or in-solution tryptic digestion, aiming at the recovery of genuine PM proteins according to their physicochemical properties (Figure 21.2).

From the Less to the More Hydrophobic Proteins. In a few studies, because after two-phase partitioning PM vesicles are retrieved equally in two configurations, cytoplasmic side-in and side-out, PM vesicles are first treated by sonication or by the Brij58 detergent. Such pretreatments help to remove from the membrane vesicles peripheral proteins from the inner surface or soluble proteins enclosed inside the vesicles, respectively [12, 15]. Complexity of the membrane fractions to be analyzed can be reduced using different strategies and several methods have been developed to analyze membrane proteins from the less to the most hydrophobic ones [10, 16–18]. Regardless of targeted strategies, one can use differential extraction protocols to retrieve proteins with diverse level of hydrophobicity.

Salt (KCl, NaCl, KNO₃, etc.) treatments were demonstrated to abolish electrostatic interactions of peripheral membrane proteins with integral membrane proteins or the polar head of lipids. KCl [12] and NaCl [13] were thus used to wash *A. thaliana* PM fractions. As a consequence, many peripheral proteins are eliminated after these types of treatments, when most of the integral membrane proteins are still embedded in the vesicle bilayers.

Alkaline (NaOH, Na₂CO₃) treatments of membrane vesicles are expected to eliminate not only peripheral membrane proteins but also some lipid-anchored proteins and proteins that interact through hydrophobic interactions with integral membrane proteins or the polar head of lipids, especially with NaOH treatment [10] which appeared to be more drastic than NaCl treatment [16]. As a consequence, many proteins are lost after these types of treatments and essentially true integral membrane proteins remain attached to the vesicle bilayers. The use of membrane washes

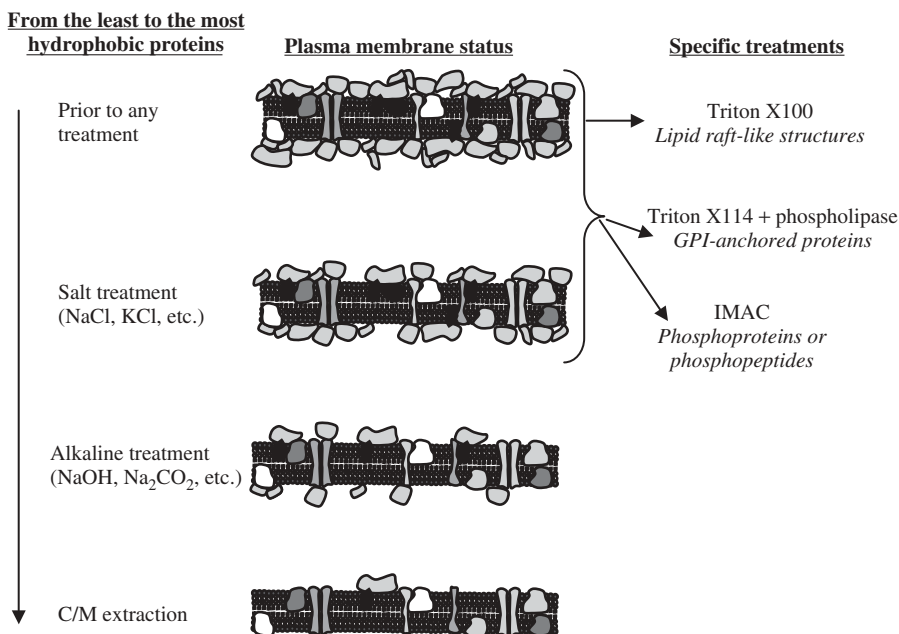


FIGURE 21.2. Outline of complementary strategies for the specific recovery of PM proteins. The left part of the figure indicates strategies aiming at the recovery of the largest proteome. They allow isolating proteomes differing at least partly by the physicochemical properties of the identified proteins. The compilation of these proteomes led to the more extensive overview of the PM proteome so far. On the right part of the figure, different strategies devoted to the retrieve specific protein categories are listed. See insert for color representation of this figure.

with NaOH [10] and especially carbonate buffers (Na_2CO_3 or $(\text{NH}_4)_2\text{CO}_3$) was frequently used for the proteome analysis of plant PMs [10, 12–15, 19, 20]. A precise analysis of this set of PM proteomes highlights the high efficiency of carbonate buffers for isolating a large variety of proteins in terms of physicochemical and functional properties.

Treatments of membrane vesicles by organic solvents, like C/M, are expected to extract the most hydrophobic proteins of the analyzed membrane system. It was originally demonstrated that most of the proteins that are extracted in organic solvents contain a high ratio of residues number present in TMDs [21, 22]. This method was applied to *A. thaliana* PM fractions [10] in order to retrieve the most hydrophobic proteins of these fractions. Although the yield of proteins extracted from a purified membrane fraction is low (5%) [8], C/M extraction allows highly hydrophobic proteins to be enriched and, thus, further identified via a proteomics approach.

Salt and alkaline treatments, as well as organic solvent extraction and detergent use, are expected to solubilize proteins from the most hydrophilic to the more hydrophobic ones. Other specific treatments, as shown in the following section, can be applied to PM fraction to extract proteins showing specific features.

Targeted Extractions (GPI-Anchored, Lipid Rafts, Phosphorylated Proteins)

Targeted extraction procedures have been described to recover and identify specific proteins. Indeed, due to its specific location between the cytosol and the extracellular medium, the PM contains proteins that have physicochemical properties that are directly linked to a specific function within the PM: GPI-anchored proteins, proteins present in lipid rafts, and phosphorylated proteins (Figure 21.2).

GPI-Anchored Proteins. GPI proteins are soluble proteins that bear a glycosylphosphatidylinositol moiety. They are mainly targeted to the PM; and because most of them lack a TMD and have no cytoplasmic tail, they are located exclusively on the extracellular side of the PM. Indeed all GPI-anchored proteins are initially synthesized with a TM anchor, but after translocation across the membrane of the ER, the ecto-domain of the protein is cleaved and covalently linked to a preformed GPI anchor by a specific transamidase enzyme. GPI-anchored proteins were shown to be enriched in lipid raft microdomains [23], and they form a diverse family of molecules that includes membrane-associated enzymes, adhesion molecules, activation antigens, differentiation markers, proteins involved in cell–cell recognition and so on. In the context of plant biology, they play a key role in pollen germination and tube growth [24] and cell wall remodeling [25].

GPI-anchored proteins can be released from membranes with specific phospholipases and can be recovered from the hydrophilic phase obtained after Triton X-114 treatment of membranes. From a PM fraction of *Arabidopsis*, the specific recovery of GPI-anchored proteins was achieved using a treatment with phosphatidylinositol PLC [20, 26] or PLD [27]. Proteomic identification of GPI-anchored proteins can be complemented by *in silico* approaches. Indeed, tools such as big-PI (http://mendel.imp.ac.at/gpi/plant_server.html) or DGPI (http://129.194.185.165/dgpi/index_en.html) provide computational methods for the assignment of GPI-anchored proteins. For instance, Elortza et al. [20, 27] showed that most of the proteins assigned as being GPI-anchored by a direct proteomics study were actually predicted by DGPI to be GPI-anchored. This suggests that such tools show low false-positive rates and can be used to predict a potential GPI-anchor for PM proteins.

Proteins Located in Lipid Rafts (Detergent-Resistant Microdomains).

Lipid rafts are membrane microdomains enriched in cholesterol or cholesterol-like (in plants) lipids. The original concept of rafts was used as an explanation for the transport of cholesterol and diverse proteins from the Golgi network to the PM. In animal cells and in yeasts, lipid rafts have been also shown to be implicated in cell-surface processes such as signal transduction, pathogen entry, or secretion. Proteins that have been shown to be associated to the lipid rafts include GPI-anchored proteins, doubly acylated tyrosine kinases of the Src family, and TM proteins. In plant cells, such lipid rafts have been pointed out [reviewed in reference 28] and studied by proteomics approach. Like in animal cells, the plant lipid rafts are enriched in sterols and sphingolipids that form a liquid ordered phase inside the membrane, which is resistant to non-ionic detergents. The extraction of lipid raft proteins would take advantage of

lipid raft resistance to non-ionic detergents, such as Triton X-100 or Brij-98 at low temperatures (e.g., 4°C). When such a detergent is added to cells, the fluid membrane will dissolve while the lipid rafts may remain intact and could be extracted using a subsequent sucrose gradient. The lipid raft fractions are called detergent-resistant membranes (DRMs). Alongside the determination of the lipid content of those lipid rafts, proteomic analyses took advantage of the Triton X-100 based protocol to identify proteins that are located in plant PM lipid rafts [5, 6, 23]. In particular, Morel et al. [6] described an extensive proteomic study of plant PM DRMs and could identify 145 proteins potentially associated to DRMs. In that study, the 145 DRMs proteins were compared to 330 proteins identified from plant PM proteomes at that date [6]. From that comparison, no significant different physicochemical properties (pI, Gravy index, number of TMDs, and protein MW) could be cited between DRMs proteins and whole PM proteins. With regard to the PTMs borne by DRMs proteins, both of them [6, 23] identified proteins that are potentially or known to be GPI (e.g., SKU5, fasciclin-like arabinogalactan), palmitoylated (C 16:0 fatty acid), or myristoylated (C 14:0 fatty acid) proteins. This supports the notion that, like in animal or yeast cells, some proteins bearing a specific lipid anchor (GPI, palmitate or myristoylate) are targeted to plant PM DRMs [29, 30]. The functional classification of the proteins retrieved in DRMs will be treated in the next part of the text.

Phosphorylated Proteins. In eukaryotic cells, protein phosphorylation is probably the most important regulatory event. Many enzymes and receptors are switched “on” and “off” by phosphorylation and dephosphorylation. Being the sensor of the cell, the phosphorylation status of PM proteins, such as RLKs, can provide valuable clues for the understanding of signaling processes. Many PM proteins are known and/or are very likely to be phosphorylated according to the cell status (e.g., stress). For instance, the phosphorylation of some plant PIP proteins was shown to occur in response to increasing water potential [31]. Using targeted methodological strategies, proteomics can give an insight over the phosphorylation status of proteins. In the context of the plant, more than 300 phosphorylation sites were identified in the PM proteome corresponding to more than 200 proteins [15]. In that study, PM fractions were digested by trypsin and phosphorylated peptides were enriched using IMAC columns, for subsequent LC–MS/MS experiments. New phosphorylation sites were pointed out, especially on RLKs, which revealed an unexpected complexity of regulation possibilities.

MS-Based Analysis for Protein Identification

Because PM contains hydrophobic proteins SDS-PAGE is highly preferred to 2-DGE. Consequently SDS-PAGE separation is the preferred electrophoretic technique to handle plant membrane proteins (Table 21.1). Proteins can be separated in the separating gel or concentrated between the separating and the stacking gels, prior to classical trypsin in-gel digestion [10, 16]. Alternatively, membrane proteins can be digested in 60% methanol [32], but this was not described in the case of PM membrane proteins. In-solution trypsin digestion was used, using a classical buffer [33].

After membrane proteins are separated by SDS-PAGE, (2D) LC–MS/MS analyses are generally carried out.

21.3 EXPERIMENTAL RESULTS AND APPLICATIONS: TOWARD AN INFORMATIVE DATABASE OF THE PLANT PM PROTEOME

As quoted above, complementary proteomic approaches at different levels (material, PM preparation, extraction, analytical tools, and targeted analyses) were used to identify proteins present in plant PM fractions. A first tentative plant PM database was described few years ago but was mainly based on 2D gel experiments [34]. Subsequently, many proteomic approaches (Table 21.1) have allowed the identification of numerous PM proteins to be carried out. Considering those recent reports, almost 1000 (989) proteins, after removal of obvious contaminants (ribosomal proteins and proteins from other cell compartments), were identified as known or putative plant PM proteins. Those proteins can be found in the website <http://www.grenoble.prabi.fr/data/PlantProteomics07> in a document where the following are compiled: (i) description and functional annotations, (ii) results of prediction bioinformatics tools (TM domains, cell localization, etc.), (iii) information (function, localization, etc.) retrieved from different protein databases (TAIR, NCBI, and ExPaSy), and (iv) literature mining data.

Hydrophobic or Lipid-Anchored PM Proteins

Schematically, two types of membrane proteins can be found in the PM. First, integral proteins go across and strongly interact with the membrane usually through at least one hydrophobic TM α -helix or less frequently with a β -barrel structure. Peripheral proteins are a second class of membrane protein. They do not have typical hydrophobic domains but interact with the membrane via amphipathic helices, hydrophobic loops, covalent links to lipid anchors, electrostatic interactions with the lipids of the membrane, or even PPIs. Because many hydrophobic proteins are present in membranes, prediction tools can be used to predict proteins that have potentially α -helical domains [35]. Thus, according to the Aramemnon database, about 40% of the compiled PM proteome can be considered as composed by integral proteins, because they possess at least one TMD (Figure 21.3).

With respect to peripheral proteins, PTMs by lipid moieties is known as an important clue allowing dedicated targeting of a protein to a membrane via anchoring into the lipid bilayer. However, lipidation of a protein remains extremely difficult to identify by protein chemistry-based methods followed by MS analyses. This is due to a number of reasons including, for instance, (i) instability at the sample fractionation or MS/MS stages, (ii) strong hydrophobicity of the tryptic peptides inducing loss at both the solubilization/recovery step and the LC separation by, and (iii) complex modification patterns involving both proteolytic and acylation, making MS database searches challenging to say the least. In this context, lipid modifications are usually only predicted using dedicated software tools such as TermiNator or DGPI. From the compiled PM proteome, 17% (165/989) of the proteins were predicted to be modified by either

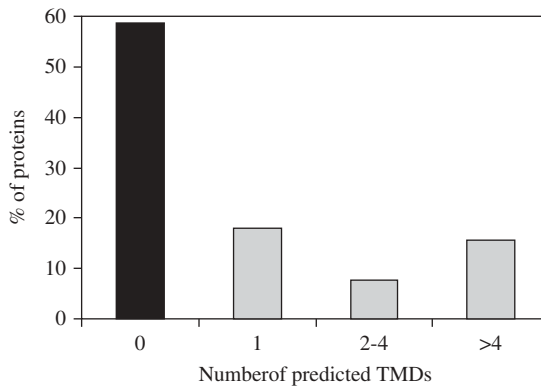


FIGURE 21.3. Distribution of the number of TMDs predicted for each protein of the compiled PM proteome. Predictions of membrane-spanning regions (i.e., TMD) were collected from the ARAMEMNON database [35] (<http://aramemnon.botanik.uni-koeln.de/>)

GPI-anchors or *N*-myristoyl, *S*-palmitoyl, or *C*-prenyl moieties (Figure 21.4). Recent work in yeast has established that palmitoylation could occur not only next to a myristoylation or prenylation site but also on a cytosolic-oriented cysteine located next to a TM domain [36]. Such modification, whose prediction is not yet feasible, could significantly increase the hydrophobicity of proteins with a single TM domain and stabilize anchoring to the membrane as a result. In most cases, however, the lipidated proteins that were annotated do not have any predicted TMD but rather display double acylation (85% of the total). Double acylation is believed to make membrane binding very stable [37]. For mono-lipidated proteins, membrane binding is likely to involve a second signal such as electrostatic interactions or PPI domains.

Therefore, when taking into account the almost 1000 proteins retrieved from plant PM proteomes, almost 60% of them are potentially embedded in the PM through either α -helices or lipidic anchors.

A Functional Survey of PM Proteins

A first remark concerns the predicted subcellular localization of the whole set of proteins compiled from the proteomes listed in Table 21.1 (989 proteins). Basically, considering Aramemnon predictions, mitochondria- and chloroplast-targeted proteins represent 7% each of the total. The rest is divided into two categories, 60% with no predicted targeting and 26% (260 proteins) predicted as secreted proteins, strongly suggesting a PM association. Altogether, the great majority of the proteins would be genuine PM proteins, more or less tightly associated with the membrane as integral, peripheral, or PTM-anchored proteins. Two arguments are supporting this conclusion: the redundancy between the different proteomes (many proteins are identified in several PM proteomes) and the functional survey of the compiled proteomes. Indeed, the functional proteome is fully keeping in with functions known or expected from the PM. The proteins with predicted functions have been divided into seven functional

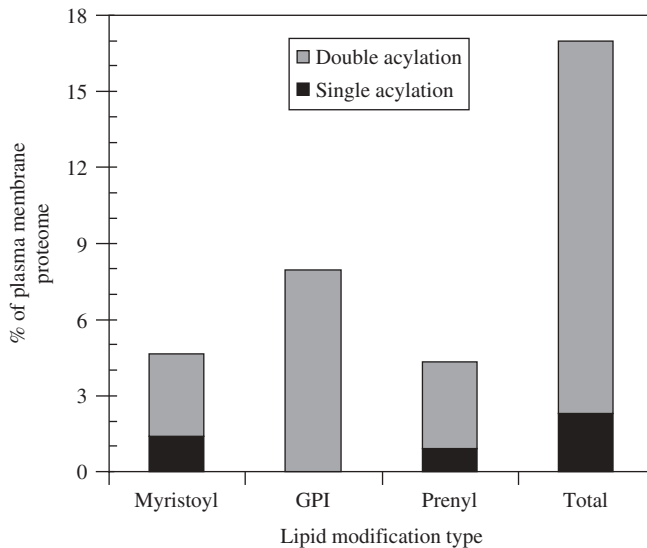


FIGURE 21.4. Impact of protein lipidation on the PM proteome of *Arabidopsis*. PTM predictions were performed on-line using various dedicated programs. N-terminal myristoylation was first predicted by Terminoator (<http://www.isv.cnrs-gif.fr/terminator2/index.html>) using relaxed criteria. Positive entries were next validated using the updated (December 2006) version of the *Arabidopsis* database of myristoylated proteins (<http://www.isv.cnrs-gif.fr/tm/maturation/myristoylome2006.html>) using the software developed by Boisson et al. [46]. S-palmitoylation on vicinal cysteines next to the N-terminus of N-myristoylated proteins was predicted with Terminoator [47]. C-terminal prenylation was predicted by PrePS (<http://mendel.imp.ac.at/sat/PrePS/index.html>). Robust predictions of GPI anchors were retrieved from the ARAMEMNON database that compiles three prediction programs (big-PI, DGPI and GPI-SOM). GPI-anchored proteins were also predicted with DGPI (http://129.194.185.165/dgpi/index_en.html).

classes, from the most to the less abundant: cell signaling, metabolism, transport, cell traffic, cell structure, protein maturation, and protein turnover (Figure 21.5). This classification is fully in agreement with the ones previously reported, although they were partial. In this part, we choose to highlight the main specific and/or novel features displayed by PM proteome (for a detailed inventory, see Table 21.1, and references herein).

An Obvious Role in Transport. The role of PM in ion, water, and solute transport is well established. Proteins involved in membrane transport represent 15% of the total PM proteome. Actually, canonical transport proteins are routinely identified, and 13 P-type H^+ -ATPases and most of the PM aquaporins (PIPs), mono- and di-saccharide translocators (10 proteins), and phosphate transporters (four proteins) are also present. Among the less well characterized transport proteins are 19 members from the large family of ABC transporters and 18 V-type H^+ -ATPase subunits [10–13]. Interestingly, a comparison between the DRMs to the whole PM protein

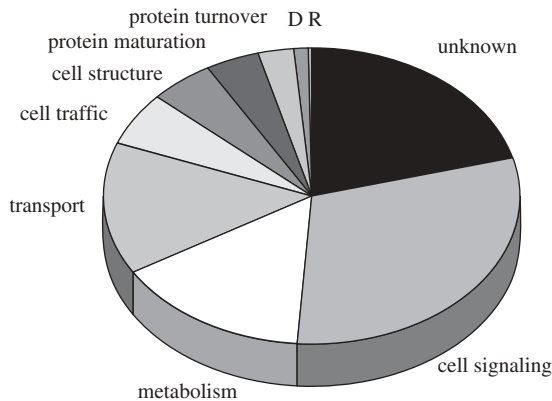


FIGURE 21.5. Distribution of the functions for the compiled PM proteome. Protein names and/or associated functions were retrieved from NCBI (<http://www.ncbi.nlm.nih.gov/entrez>), PPDB (<http://ppdb.tc.cornell.edu/>), or TAIR (www.arabidopsis.org). The whole set of proteins (989) was classified in seven functional categories; few proteins were grouped in two other classes, “D” for DNA structure (nine proteins) and “R” for RNA transcription (eight proteins), which more likely are contaminants. The unknown category in the “camembert” concerns proteins with no annotated function

content indicated that ATPases and aquaporins tend to be enriched in DRMs [5, 23]. Transport proteins are likely to be deeply embedded in the membrane, and most of them are likely to be hydrophobic proteins; for this reason, this class is maybe the most underrepresented in the proteome, because technical difficulties have still to be overcome to retrieve proteins displaying such physicochemical properties.

A Seminal Role in Signaling. The main protein category of the compiled PM proteome is devoted to cell signaling (31.5%). Besides kinases (2.7%)/phosphatases (1.6%), calcium-binding proteins (2.1%) (such as CaMs and CBPs), interacting/regulatory proteins (4.4%) (such as remorins, 14–3–3 isoforms, regulators of cell cycle like CDC48A, CYCB1), and proteins for cell detoxication in response to biotic or abiotic stress (3.4%), the most abundant subclasses are GTP-binding proteins (45/989, 4.6%) and proteins involved in signal perception (118/989, 11.9%). Most of the identified G-proteins have no assigned biological function in signal transduction pathways. Nevertheless, the large number of small GTP-binding proteins from the large RAB family is in good agreement with their role in delivering of new cell-wall components to the PM and, more generally, to the regulation of endocytic trafficking pathway [38]. As to the signal perception components, our knowledge is somehow better. Some have already been characterized and their perceived signal is known; this is the case for Mlo, a seven-TMD protein conferring genetically mildew resistance [39], the two members of the brassinosteroid (BR) receptor complex, BRI1/BAK1 [40], and ARK3, an S-locus kinase involved in self-incompatibility [41]. But the large majority of the receptors putatively located in the PM remains unknown; it includes most of the 105 RLK proteins present in the compiled proteome.

It is noteworthy that proteins involved in plant defense (e.g., remorin, NADPH oxidase, small G protein Ntrac5, HR proteins, etc.) were enriched in DRMs fractions [5, 23].

A Membrane in Tight Interaction with Other Cell Compartments. Due to the vicinity of cell wall, the presence of proteins active at the interface of the two compartments is a typical feature displayed by PM proteome. In particular, proteins participating in the cell-wall composition and in the biosynthesis of cell-wall components are frequently identified in the PM proteomes. Several of them, such as the strictosidine synthase and the endotransglycosylase MERI-5, have also been reported in the cell-wall proteome [42], showing the faint barrier between PM and cell-wall proteins. Functionally, the presence in the PM of these enzymes and of the RAB-A G-proteins suggests that these processes are regulated, at least partly, at the PM level.

On the other side, the PM communicates with the interior of the cell. First, PM occupies a key position for interacting with the cytoskeleton, which is supported by the identification of proteins belonging to large families such as profilins and kinesins, both known to regulate cytoskeleton dynamics. Second, increasing evidence for the role of PM as a site for exchanges is brought forth by the presence in the compiled PM proteome of numerous proteins (6%) characterizing endocellular trafficking and secretion processes (SNAREs, syntaxins, dynamins, Sec proteins, etc.). These data illustrate the interplay between PM and the various endomembrane systems of the cell.

PM Functions in Emergence. The most recent proteomes [6, 13] pointed out the presence of proteins participating in protein maturation–translation (elongation factors EF-1 and -2, initiation factors EIF), folding (HSP, peptidyl-prolyl *cis*–*trans* isomerases), and protein modifications (arginine *N*-methyl transferase, UDP-glucosyl transferase). A second new category is composed of proteins involved either in protein turnover—22 proteins are participating in the Ub-proteasome pathway, such as several subunits of the proteasome, Ub-protein ligases, and Ub-conjugating enzymes—or in proteolysis (proteases, aminopeptidases, one caspase). These proteins define two new functional classes, protein maturation and protein turnover, representing 4.3% and 2.9% of the compiled proteome, respectively. These proteins are mainly soluble, but their biological role may involve transient interaction with the PM.

Still Unknown Proteins. Based on both sequence homologies with known proteins and identification of predicted conserved domains, only 18% of the total proteins are classified as unknown (Figure 21.5). It is noteworthy that this value is significantly lower than those reported in the original papers (around 25–30%), which is likely attributed to the recent progress in *Arabidopsis* genome annotation.

21.4 CONCLUSIONS

The major interest of such a subproteomics approach dealing with PM is that it presents clues on both the membrane compartment where proteins are working and their putative cellular functions. The PM is a site of very specific functions, and consequently

the PM proteome is expected to gather many different types of proteins with regard to their physicochemical properties. Therefore, as illustrated by the analysis of the compilation of almost all PM proteomes reported so far, it is necessary to combine complementary protein extraction procedures to retrieve the most exhaustive protein inventory. This analysis confirmed the wide diversity of the proteins as expected; indeed, they display physicochemical characteristics over a large range of values for pI, MW, GRAVY index, and so on; 40% have from 1 to 16 predicted TMDs and 17% display lipid modifications. The functional analysis corroborated functions classically associated with PM, such as transport, endocellular trafficking, and secretion, but it also spotlighted the unique role of PM in cell signaling.

21.5 FIVE-YEAR VIEWPOINT

A first level of validation of a PM proteome relies on the purity of the membrane fraction, which still remains a challenge for PM. Especially extraction and purification methodologies have to be improved when protein extraction is addressed to organs or specific tissues present in such small quantities in *Arabidopsis* because it constitutes a bottleneck for biochemical approaches. Indeed, this step is a necessity for progressing toward functional proteomics approaches investigating PM functions related to different physiological situations during plant growth and development and plant responses to abiotic and biotic stress.

A second major goal for the next future is to overcome the status of *in silico* predictions for the proteome by developing high-throughput methodologies for confirming, or invalidating, the PM localization of the identified proteins. *In silico* approaches have been described to predict the localization of plant membrane proteins, especially for chloroplast and mitochondrial membrane systems. For these two organelles, tools can predict the specific targeting, because many chloroplastic and mitochondrial proteins bear a signal transit peptide. This is not the case for PM proteins. Indeed, no specific signature or signal peptide targeting proteins to the PM have been identified so far. Consequently, experimental localization remains up to now the only way to get this information. A very few PM proteomics studies are supported by subcellular localization results and remain limited to the study of some targeted proteins [10, 33]. Escobar et al. [43] developed an innovative strategy based on a high-throughput viral expression system in *N. benthamiana* of a cDNA-GFP fusion library. More recently, a protein localization database has been developed by Li et al. [44] for facilitating fluorescent tagging of full-length *Arabidopsis* proteins using pre-selected GFP insertion position and primer sequences for all *Arabidopsis* proteins. No doubt that more extensive subcellular localization studies would emerge soon.

The investigation of the biological function of the identified proteins would be likely much more difficult to solve with regard to the large diversity of functions associated with the PM. Now it is time to take up the challenge to set up high-throughput methodologies for assaying protein activities. Technologies based on the use of heterologous systems for expressing proteins, as suggested for transport proteins [45], are maybe one of the clues. Though this question is especially relevant for the PM proteome, it is not specific to it and general strategies have to be developed.

Indeed, due to the specific status of the PM, which is a highly dynamic membrane system with respect to the cell type, the environment, its sensor abilities, and so on, quantitative proteomic approaches provide valuable tools to investigate further the PM proteins. In a proof-of-concept paper, Lanquar et al. [19] showed that *A. thaliana* suspension cells could undergo efficient ^{15}N labeling. This metabolic labeling was further used to find out which proteins, in a soluble and in a PM fraction, was up- or down-regulated after Cd stress. Such a quantitative proteomic strategy could be easily applied to study the effect of different stresses at the level of the PM. Moreover, being the sensor of the cell, the plant PM is the starting point for different signaling pathways. A recent PM proteomics study showed that receptor kinases are highly abundant in the PM [13]. Those data, together with the identification of the occurrence of phosphorylated proteins in the PM [15], emphasize the sensor role of the PM and the potentials of some PM proteins to trigger specific cell behavior in a bioengineering context.

REFERENCES

1. Santoni V, Molloy M, and Rabilloud T. 2000. *Electrophoresis* **21**:1054–1070.
2. Wu CC, et al. 2003. *Nat Biotechnol* **21**:532–538.
3. Santoni V, et al. 1998. *Plant J* **16**:633–641.
4. Santoni V, et al. 1999. *Biochimie* **81**:655–661.
5. Mongrand S, et al. 2004. *J Biol Chem* **279**:36277–36286.
6. Morel J, et al. 2006. *Mol Cell Proteomics* **5**:1396–1411.
7. Larsson C, Widell S, and Kjellbom P. 1987. *Methods Enzymol* **148**:558–568.
8. Marmagne A, et al. 2006. *Methods Mol Biol* **323**:403–420.
9. Santoni V. 2007. *Methods Mol Biol* **355**:93–109.
10. Marmagne A, et al. 2004. *Mol Cell Proteomics* **3**:675–691.
11. Ephritikhine G, Ferro M, and Rolland N. 2004. *Plant Physiol Biochem* **42**:943–962.
12. Alexandersson E, et al. 2004. *Plant Cell Physiol* **45**:1543–1556.
13. Marmagne A, et al. 2007. *Mol Cell Proteomics* **6**:1980–1996.
14. Nelson CJ, et al. 2006. *Mol Cell Proteomics* **5**:1382–1395.
15. Nuhse TS, et al. 2004. *Plant Cell* **16**:2394–2405.
16. Ferro M, et al. 2003. *Mol Cell Proteomics* **2**:325–345.
17. Friso G, et al. 2004. *Plant Cell* **16**:478–499.
18. Peltier J, et al. 2004. *J Biol Chem* **279**:49367–49383.
19. Lanquar V, et al. 2007. *Proteomics* **7**:750–754.
20. Elortza F, N et al. 2003. *Mol Cell Proteomics* **2**:1261–1270.
21. Seigneurin-Berny D, et al. 1999. *Plant J* **19**:217–228.
22. Ferro M, et al. 2002. *Proc Natl Acad Sci USA* **99**:11487–11492.
23. Borner GHH, et al. 2005. *Plant Physiol* **137**:104–116.
24. Lalanne E, et al. 2004. *Plant Cell* **16**:229–240.
25. Roudier F, et al. 2005. *Plant Cell* **17**:1749–1763.

26. Borner GHH, et al. 2003. *Plant Physiol* **132**:568–577.
27. Elortza F, et al. 2006. *J Proteome Res* **5**:935–943.
28. Bhat RA and Panstruga R. 2005. *Planta* **223**:5–19.
29. Zacharias DA, et al. 2002. *Science* **296**:913–916.
30. Resh MD. 1999. *Biochim Biophys Acta* **1451**:1–16.
31. Maurel C. 1997. *Annu Rev Plant Physiol Plant Mol Biol* **48**:399–429.
32. Jaquinod M, et al. 2007. *Mol Cell Proteomics* **6**:394–412.
33. Dunkley TPJ, et al. 2006. *Proc Natl Acad Sci U S A* **103**:6518–6523.
34. Sahnoun I, et al. 2000. *J Biotechnol* **78**:235–246.
35. Schwacke R, et al. 2003. *Plant Physiol* **131**:16–26.
36. Roth AF, et al. 2006. *Cell* **125**:1003–1013.
37. Resh MD. 2004. *Subcell Biochem* **37**:217–232.
38. Vernoud V, et al. 2003. *Plant Physiol* **131**:1191–1208.
39. Buschges R, et al. 1997. *Cell* **88**:695–705.
40. Nam KH, Li J. 2002. *Cell* **110**:203–212.
41. Sherman-Broyles S, et al. 2007. *Plant Cell* **19**:94–106.
42. Bayer EM, et al. 2006. *Proteomics* **6**:301–311.
43. Escobar NM, et al. 2003. *Plant Cell* **15**:1507–1523.
44. Li S, Ehrhardt DW, and Rhee SY. 2006. *Plant Physiol* **141**:527–539.
45. Miller AJ and Zhou JJ. 2000. *Biochim Biophys Acta* **1465**:343–358.
46. Boisson B, Giglione C, and Meinnel T. 2003. *J Biol Chem* **278**:43418–43429.
47. Meinnel T, Peynot P, and Giglione C. 2005. *Biochimie* **87**:701–712.

NUCLEUS

Subhra Chakraborty, Aarti Pandey, Asis Datta, and
Niranjan Chakraborty

22.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

The nucleus is common to all eukaryotic cells and houses nearly all the genetic information for the expression of cellular proteins. It contains the components and enzymes necessary to maintain, transcribe, and replicate genetic material in a selective manner. Furthermore, it is responsible for the synthesis and/or assembly of components used to translate genetic material, although translation of the code contained in mRNA is performed in the cytoplasm. The nucleus consists of a complex envelope—an outer and inner membrane that surrounds a relatively amorphous internal structure and an internal structure consisting of chromatin, nuclear matrix, nuclear lamina, nucleolus, and aqueous nucleoplasm.

The nuclear proteome is dynamic, changing its composition in response to intracellular and environmental stimuli. Proteins involved in different cellular functions—for example, signaling, gene regulation, structure, translation, proteolysis, and, among others, a variety of RNA-associated functions—have been identified in the nucleus. Increasing evidence suggests that nearly 27% of total cellular proteins are localized

in the eukaryotic nucleus, implying a variety of protein functions in this compartment [1]. Some plant components of intranuclear compartments were reported to differ greatly from those of other organisms. Only a few plant nuclear matrix proteins have been characterized, and they have no obvious homology with known nuclear proteins in yeast and mammals [2, 3]. Therefore, characterization of the nuclear proteome in plants will go a long way in increasing our understanding about the gene expression and function in a plant system.

The basic problem of complexity poses a significant challenge in studies to unravel the protein complement of the genome, the proteome. Many of the estimated genes in a genome are expected to provide multiple protein products that might arise as a result of alternate splicing and PTM. Proteomics is a leading technology for large-scale analysis of protein expression, splice variants, PTMs, and PPIs [4]. However, the resolution of protein spots on a 2D gel is limited by factors such as protein abundance, size, hydrophobicity, and other electrophoretic properties [5]. One approach to the daunting prospect of cataloging entire proteomes has been to focus on protein subsets. Organelles, the subcellular compartments in a dynamic intracellular membrane system, provide such subsets because they can be subfractionated [6]. Furthermore, the compartment-specific proteome is preferred because fractionated subsets of proteins provide the suitable information in which they exert their particular function [7]. The subcellular proteome research is quite advanced in animals, yeast, and *E. coli* [8]; however, plant subcellular proteomics is still in its infancy. The study on the nuclear proteome has been undertaken by several groups, but most efforts have been largely restricted to mammals, which include subnuclear fraction from embryos of *Drosophila* [9], nuclear matrix proteins in various human cell types [10–12], nuclear envelope proteins from mouse neuroblastoma N2a cells [13], human nucleolar proteins [14], and total nuclear proteins from human liver [15]. Recently, proteomic analyses of nucleus in *A. thaliana* [16], and as a very preliminary study in rice [17], have been reported. The proteomic analysis of the *Arabidopsis* nuclear matrix [18], as well as nucleolus [19], has also been published.

Very recently, we have developed a reference proteome map for a legume crop, chickpea [20], as a basis for future proteome comparisons of genetic mutants and pathogen-infected and/or environmentally challenged plants, with special emphasis on understanding the nucleus-specific cell-signaling network. Approximately 600 proteins were resolved using 2-DGE; 150 of these proteins were identified and classified into different functional categories. The proteins that we have identified reveal the presence of complex regulatory networks that function in this organelle. The comparison of nuclear proteomes of different plant species illustrates divergence in protein profile with only few conserved proteins. The differences in terms of protein pattern and protein function appear to encompass both genetic and physiological information.

22.2 METHODOLOGY AND STRATEGY

Plant Material

Legumes are valuable agricultural and commercial crops that serve as important nutrient sources for human diet and animal feed. About one-third of human nutrition

comes from legumes; in many developing countries, legumes serve as the only source of protein. They are essentially characterized by symbiotic relationships with both nitrogen-fixing bacteria and AM fungi [21]. Chickpea (*C. arietinum* L.) is one of the most important legume crops. The 3-week-old seedlings were sampled as experimental materials, harvested, and stored at -80°C after quick-freezing in liquid nitrogen.

Isolation of Nuclei and Protein Extraction

The nuclei were prepared from chickpea seedlings as described [20]. About 20 g of the tissue was ground into powder in liquid nitrogen with 0.3% (w/w) PVPP and immediately transferred into an ice-cold 500-mL beaker containing 200 mL of ice-cold 1X HB (10 mM Trizma base, 80 mM KCl, 10 mM EDTA, 1 mM spermidine, 1 mM spermine, 0.5 M sucrose, pH 9.5) plus 0.15% 2-ME and 0.5% Triton X-100. The contents were gently stirred for 30 min for complete lysis of organeller membranes. This suspension was filtered through four layers of cheesecloth and two layers of Mira cloth into an ice-cold 250-mL centrifuge bottle. The homogenate was pelleted by centrifugation with a fixed-angle rotor at 1800g at 4°C for 20 min. The supernatant fluid was discarded and the pellet was gently resuspended in 30 mL of ice-cold wash buffer (1X HB minus Triton X-100). To remove the particulate matter remaining in the suspension, the resuspended nuclei were filtered into a 50-mL centrifuge tube through two layers of Mira cloth by gravity. The content was centrifuged at 57g, 4°C for 2 min to remove intact cells and tissue residues. The supernatant fluid was transferred into a fresh centrifuge tube and the nuclei were pelleted by centrifugation at 1800g, 4°C for 15 min in a swinging bucket centrifuge. The pellet was washed two additional times by resuspension in wash buffer followed by centrifugation at 1800g, 4°C for 15 min. Nuclear proteins were prepared from the nuclei-enriched pellet using TriPure Reagent (Roche) according to the manufacturer's instructions. The final protein pellet was resuspended in IEF sample buffer [8 M urea, 2 M thiourea, 2% (w/v) CHAPS]. The protein concentration was determined using the 2-D Quant kit (Amersham Biosciences).

Purity Assessment of the Isolated Nuclear Fraction

The isolated nuclear fraction was checked for contamination at two levels: the organellar and the protein level. At the organellar level, the nuclear pellet was subjected to microscopic analysis as well as checked for chlorophyll contamination. The isolated nuclei were visualized under a phase-contrast microscope for any cross-contamination from other organelles. The chlorophyll content in the starting homogenate, the supernatant, and the nuclei-enriched fraction was determined using a spectrophotometric assay. The sample was prepared by pipetting 1 mL of suspension into a 15-mL centrifuge tube and adding 8 mL of acetone and 1 mL of MQ water to it and centrifuging at 1000g for 5 min. The absorbance of this sample was measured at 652 nm. The assay was done in triplicates and the amount of chlorophyll in 1 mL of the suspension was observed as $\text{mg chlorophyll/mL} = \text{absorbance}/34.5$ (<http://www.bio.com/protocolstools/protocol.html>). The chlorophyll amount was then calculated as milligram

per gram of fresh tissue weight. The purity of the nuclear fraction was evaluated on the basis of the difference in chlorophyll content in supernatant and the nuclear suspension.

At the protein level, enrichment was checked by immunoblotting for nuclear resident proteins and by catalase assay. For immunoblotting, proteins were separated on 12.5% SDS-PAGE and electrotransferred onto nitrocellulose membranes. Membranes were incubated with mouse anti-fibrillarin and sheep anti-histone antibodies (Abcam Limited, UK). Antibody-bound proteins were detected by incubation with secondary antibodies (Abcam) conjugated to AP. The catalase enzyme assay was performed using 10 µg of organellar protein for each reaction. The reaction mixture was prepared by adding 50 µL of protein extract to 940 µL of 70 mM potassium phosphate buffer (pH 7.5). Reaction was started by addition of 10 µL of H₂O₂ (3% v/v), and the decrease in absorbance at 240 nm was followed for 5 min. Baseline correction was done by subtracting the absorbance taken without addition of H₂O₂. The assay was done in triplicates, and the absorbance values obtained were plotted against time (<http://www.hos.ufl.edu/meteng/HansonWebpagecontents/Cellfractionation.html>).

2-DGE and Identification of Nuclear Proteins

IEF was carried out with 150 µg of protein. Aliquots of proteins were diluted with 2D rehydration buffer [8 M urea, 2 M thiourea, 2% (w/v) CHAPS, 20 mM DTT, 0.5% (v/v) pharmlayte (pH either 3–10, 4–7, or 6–11) and 0.05% (w/v) BPB], and 250 µL of solution was used to rehydrate IPG strips (13 cm; pH 3–10 and 4–7). Protein was loaded by in-gel rehydration method onto IEF strips, and electrofocusing was performed using IPGphor system (Amersham Biosciences, Bucks, UK) at 20°C for 30,000 V-h. However, for 6–11 pH strips, anodic cup-loading was performed with a load of 100 µg of protein in 100 µL of rehydration volume, and electrofocusing was performed for 70,000 V-h. The focused strips were subjected to reduction with 1% (w/v) DTT in 10 mL of equilibration buffer [6 M urea, 50 mM Tris-HCl (pH 8.8), 30% (v/v) glycerol and 2% (w/v) SDS], followed by alkylation with 2.5% (w/v) iodoacetamide in the same buffer. The strips were then loaded on top of 12.5% PAGE for SDS-PAGE. The electrophoresed proteins were stained with silver stain plus kit (Bio-Rad, CA). Gel images were digitized with a Bio-Rad FluorS equipped with a 12-bit camera. The PD Quest version 7.2.0 (Bio-Rad) was used to assemble a first-level matchset (master image) from three replicate 2D gels. Protein spots were excised mechanically using pipette tips and in-gel digested with trypsin (Sigma, St. Louis) according to standard techniques. Peptides were analyzed by ESI-LC-TOF-MS using a Q-Star Pulsar *i* (Applied Biosystems). The spectra were analyzed using Mascot sequence matching software (www.matrixscience.com) and the NCBI nr database with Viridiplantae (green plants) taxonomy.

22.3 EXPERIMENTAL RESULTS AND DISCUSSION

Isolation of Purified Nuclei

The most important criterion for organellar proteomics is the purity of the compartment to be analyzed. Indeed, the integrity of a subcellular proteome is largely dependent on

the purification of the isolated compartment away from other cellular contaminants. The separation of high-purity nuclei from plant is a difficult task, because it might compromise the yield. In this study, the nuclei were isolated from chickpea seedling using hyperosmotic sucrose buffer, and the nuclei-enriched pellet so obtained was washed repeatedly to get rid of contaminants from other organelles. The integrity of the isolated nuclei was analyzed under a phase-contrast microscope (Figure 22.1A). The chickpea nuclei were uniform spheres with an average diameter of approximately 20 μM , while the presence of no other organelle was observed. These results indicate that the isolated nuclei were highly purified. Possible chloroplast contamination in the nuclear fraction was examined by spectrophotometric analyses of chlorophyll.

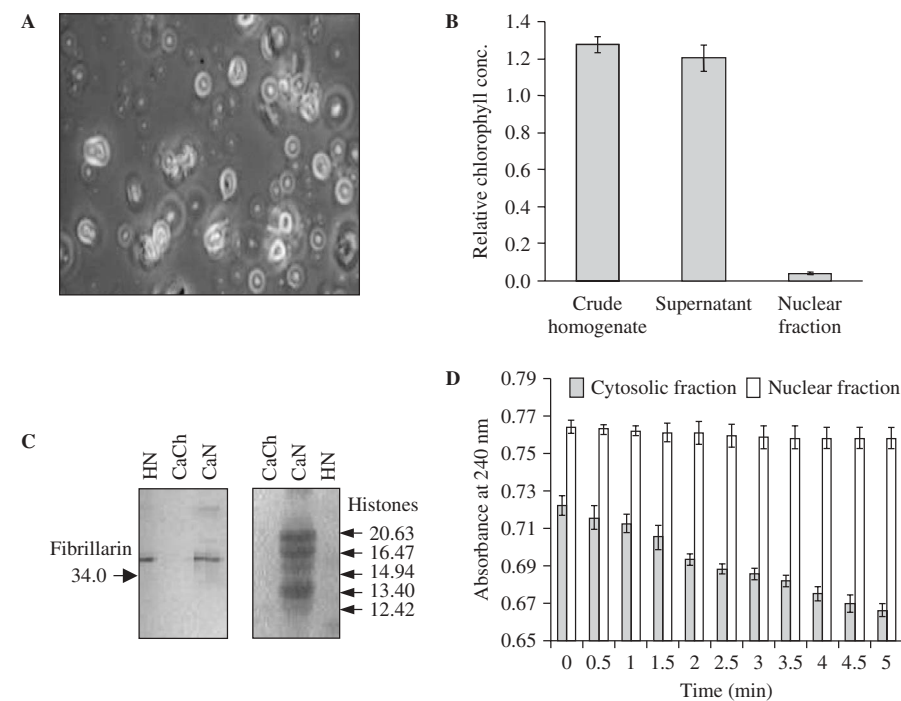


FIGURE 22.1. Purity assessment of the isolated nuclear fraction. **(A)** The purified nuclei were viewed by phase contrast microscopy to check for cross-contamination from other organelles. **(B)** Determination of chlorophyll content at different stages of purification. The amounts of chlorophyll present in crude homogenate, supernatant, and nuclear fraction were estimated and compared. **(C)** Immunoblot analysis of nuclear extracts with nuclear-resident proteins, histone and fibrillarlin. One hundred micrograms of protein each from chickpea nuclear (CaN) and chloroplast (CaCh)-enriched fractions and from HeLa nuclear extract (HN) were separated by 12.5% SDS-PAGE. The 1D gel was electroblotted onto Hybond-C membrane, and histones and fibrillarlin were detected. The MW (kDa) of the resident proteins is indicated by arrows. **(D)** Contamination from non-nuclear proteins was monitored by measuring the activity of catalase enzyme. Cytosolic fraction was used as positive control.

As shown in Figure 22.1B, the supernatant retained most of the chlorophyll content, and less than 3% chlorophyll was present in the nuclei pellet.

The nuclear proteins were prepared from the purified nuclei using TriPure reagent, in order to remove the contaminating nucleic acids that might interfere during the IEF process. The enrichment of nuclear proteins was evaluated by immunoblot analysis using specific antibodies for two nuclear proteins, histone core and fibrillarin. The nuclear resident proteins histone and fibrillarin were detected in the nuclear fraction, but not in the chloroplast fraction (Figure 22.1C). Contamination with non-nuclear proteins was monitored by measuring the activity of catalase enzyme as an indirect cytosolic marker in the nuclear fraction. While the cytosolic proteins showed high catalase activity, nuclear proteins did not show any significant catalase activity (Figure 22.1D). These results, altogether, suggest that the isolated nuclear fraction was more or less pure and free from other cross-contaminants.

Construction of 2D Map

Nuclear proteins were separated by 2-DGE in order to establish a reference map. The images were analyzed by the PD Quest software. Computational analysis of the silver-nitrate-stained gels reproducibly revealed 312 different spots in the pH range 3–10. However, proteins in the basic pH range exhibited poor resolution. To make the reference map more comprehensive, we developed the proteome in the overlapping pH ranges 4–7 and 6–11 (Figures 22.2A and 22.2B). Indeed, reproducibility of high-resolution 2D patterns is an issue of concern for the basic pH range. Thus, sample cup-loading instead of in-gel rehydration was necessary. Consequently, 482 and 361 spots were detected at pH range of 4–7 and 6–11, respectively, that included the overlapping region. More than 600 exclusive spots were detected, out of which 572 spots survived the filtering process. A total of 160 spots with more than 30 quality score assigned by the software (based on the quality and quantity of the spots) were excised. Of these, 150 spots were identified with high confidence, and the corresponding protein spots are indicated on the gels (Figures 22.2A and 22.2B).

Functional Classification of Nuclear Proteins

The identified proteins were sorted into various categories (Figure 22.3), according to their putative functions. However, the classification of proteins is only tentative because the biological function of many proteins identified has not yet been established experimentally. It is known that the same protein may have different functions in different subcellular compartments and can act as “moonlighting proteins,” as is the case with the glycolytic pathway enzymes, GAPDH and PGK [22]. In a number of cases, the same protein was found in multiple spots in the same gel, suggesting possible alternate PTMs. Interestingly, the relative positions of a single protein on the 2D gel indicated that the modifications affected the pI, the MW, or both. About 11% of the identified proteins were grouped under the unknown category because no information as to their potential function in the organelle was available. Other functional categories identified are explained below.

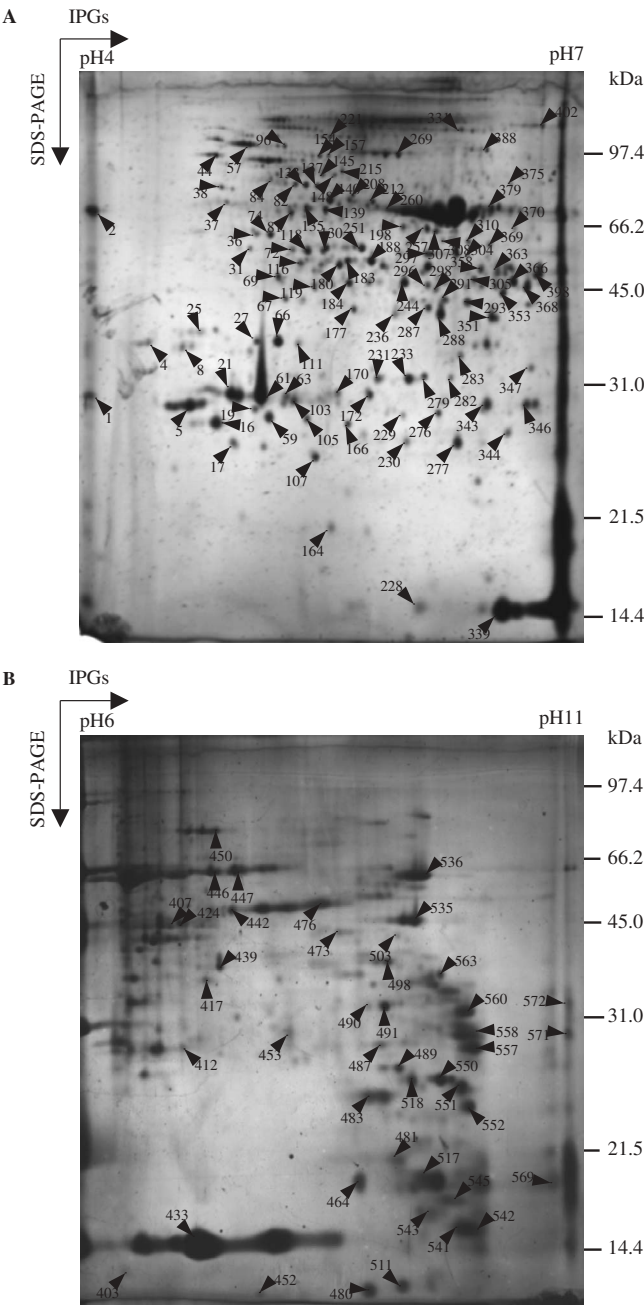


FIGURE 22.2. Nuclear proteome map of chickpea. The nuclear proteins were resolved onto (A) pH 4–7 with 150 μ g of protein and (B) pH 6–11 with 100 μ g of protein. IEF was performed on 13-cm IPG strips before being separated by SDS-PAGE on 12.5% gel. Protein spots marked by arrows were identified by LC-ESI-MS/MS.

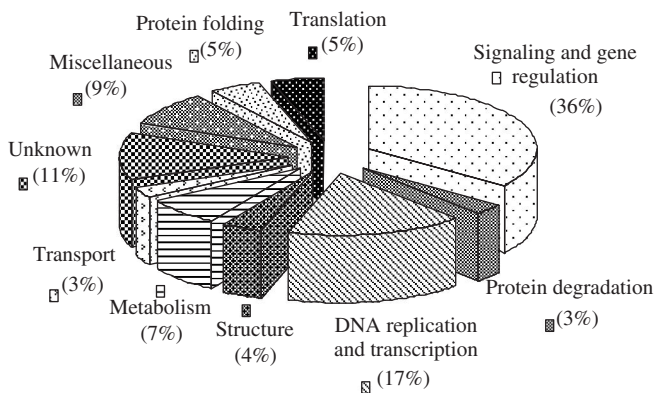


FIGURE 22.3. Functional classification of chickpea nuclear proteins. The MS-identified proteins were distributed among different protein classes based on their putative function in the organelle.

The proteins involved in signaling and gene regulation (36%) were found to be the most abundant. The proteins in this class include FCA, an RNA-binding protein. It is a plant-specific nuclear protein that is involved in flowering time control as well as ABA signaling [23]. The glycolytic enzymes, PGK and aldolase, were present as multiple spots in the chickpea nuclear proteome. These proteins are characteristic nuclear proteins in most or all eukaryotes [22]. They may have important secondary roles not directly related to carbon metabolism. Because they are known to contain a nucleotide-binding site, they may therefore bind to DNA or RNA in the nucleus. Kinases are important signaling components in any proteome—for example, Ser/Thr kinase. It is known that the majority of 14–3–3 molecules are present in the cytoplasm; however, in the absence of bound ligands, 14–3–3 homes to the nucleus [24]. The TFs are important components in any nuclear proteome, although many of these would have escaped detection because the expression level for TFs is generally not very high. Nucleocytoplasmic shuttling in turn is necessary to ensure that all these components are available at the right time in the life of a cell, and RAN-binding protein (Ras-related protein in the nucleus) acts as a major regulator of this process [25]. It is interesting to note that proteins involved in signaling and gene regulation dominated other categories, reflecting the role of nucleus in gene expression and regulation. In other organelles, like chloroplast and mitochondria, the largest percentage of proteins were reported to be involved in energy production, either in electron transport or in ATP production [26].

The second-largest category comprised proteins involved in DNA replication and transcription. Glycine-rich RNA-binding proteins (GRPs) are the predominant proteins in this category. These proteins contain two distinct domains: an amino-terminal RNA-binding domain and a Gly-rich carboxy-terminal domain. These proteins were also found to be a part of the *Arabidopsis* nucleolar proteome [19]. GAPDH is a known nuclear-resident protein. Besides its glycolytic activity, GAPDH is a multi-functional protein with both cytoplasmic and nuclear functions, one of which is as

a tRNA-binding protein participating in RNA export [27]. Asp carbamoyltransferase (ATCase) is a protein involved in *de novo* pyrimidine biosynthesis pathway. The ATCase activity is virtually absent from “isotonic nuclei” but present in nuclei isolated in hyperosmotic sucrose media [28].

Structural proteins represent the third set of nuclear proteins. These proteins are known to be major constituents of the cytoskeleton in eukaryotic cells and are involved in chromatin remodeling and related processes [29–31]. Actin is not only a major cytoskeletal component in all eukaryotic cells but also a nuclear protein that plays a role in gene transcription [32]. Histones are the chief protein components of chromatin. They act as spools around which DNA winds, enabling the compaction necessary to fit the large genomes of eukaryotes inside cell nuclei. Matrix or scaffold attachment regions (MARs or SARs) are involved in looping of open chromatin fibers. The interaction of chromatin with the nuclear matrix via MARs on the DNA is considered to be of fundamental importance for higher-order chromatin organization and regulation of gene expression. MAR-binding filament-like protein 1 (MFP1) with a filament protein-like structure is a good candidate for a MAR-binding constituent of the nuclear filaments [33].

Proteins involved in translational machinery are standard in case of any nuclear proteome. In our study, 5% of the total identified proteins belong to this category. Eukaryotic EF-1 α plays a pivotal role in protein biosynthesis, present mainly in the cytoplasm, but a small population of eEF-1 α molecules has been previously identified in the nucleus, where it forms a complex with zinc finger protein [34]. The molecular chaperones account for 5% of the total nuclear proteome of chickpea; this includes chaperonin 60 and DnaK-type molecular chaperone (HSP70). Under normal circumstances, HSP70 is present mainly in the cytosol, but it translocates to the nucleus and nucleolus during physiological stress to prevent random aggregation of proteins [35]. Another important category of proteins identified are presumably involved in degradation mechanism. Several metabolism-related proteins have also been identified across all the plant nuclear proteomes, apart from the proteins involved in transport. The miscellaneous class of proteins makes up the rest of the chickpea nuclear proteome.

22.4 CONCLUSIONS

The first and foremost requirement for a subcellular proteome lies in isolating a particular organelle away from other subcellular compartments. The purity of the isolated fraction thus becomes the most important criteria for providing direct evidence for the existence of proteins in one or more compartments within the cell. In this study, nuclear preparations consisted of homogeneous membrane-bound forms with intact nucleoli. As biochemical markers for enrichment of nuclear protein and to assess the degree of cytoplasmic contamination, immunoblot analysis and enzymatic assay were performed on nuclear fraction. The nuclear proteome of chickpea were further compared with that of *Arabidopsis* [16] and rice [17]. As expected, proteins involved in signaling and gene regulation were found to be the most abundant across the species. However, a large number of proteins were unique or novel to each of the species

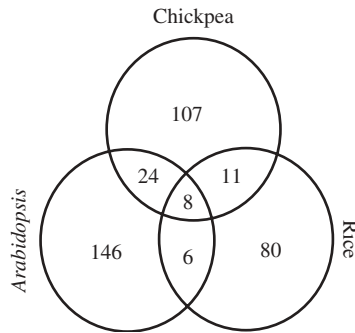


FIGURE 22.4. Comparative analyses of plant nuclear proteome. The identified nuclear proteomes of *Arabidopsis*, rice, and chickpea were compared and the overlaps depicted by a Venn diagram. The numbers signify the unique and/or orthologous proteins among the organisms studied.

studied (Figure 22.4). In conclusion, a combined approach of subcellular isolation followed by careful purification can significantly enhance the systematic identification of nuclear proteins. This information, coupled with analysis of biological processes, molecular functions, and regulatory networks, can be used to gain insights into the complexity of functions controlled by the protein machinery in the nucleus.

22.5 FIVE-YEAR VIEWPOINT

This study is directed toward the systematic analysis of the nuclear proteome in a leguminous plant, chickpea, and possible functional classification of the identified proteins. The findings may be used in the future to dissect biochemical pathways encompassed by the nucleus-specific proteins. This will also be important in long-term efforts to develop faithful, quantitative models for plant processes [36]. For a complete understanding of cellular functions, it is important to determine and characterize (a) the proteomes at different subcellular locations and (b) their involvement in biosynthetic and signaling pathways. Most of the identified proteins are verified nuclear proteins as evident by the literature. However, few of the nonclassical proteins were identified in the chickpea nucleus, which have never been associated with this compartment. These results, in part, are in agreement with the previous reports on plant nuclear proteomes, and also there is a great level of divergence in the protein classes. Nevertheless, until a much more complete survey of the proteomes of nucleus in several plants is conducted using more similar protein arraying and identification technology, it will be difficult to determine the presence/or absence of specific proteins between plant species. This is an initial attempt in the direction that will be expanded upon during future proteomics studies of plant nucleus. Our future efforts will focus on increasing the number of analyzed proteins with an aim to draw a complete functional map of nuclear proteome. Furthermore, we will focus on identifying the dynamics associated

with the nuclear proteome toward cell metabolic and regulatory pathways at different physiological conditions.

ACKNOWLEDGMENTS

This work was supported by a grant (BT/PR/4016/Agr/16/327) from the Department of Biotechnology (DBT), Government of India and a grant from National Centre for Plant Genome Research, New Delhi, India to N.C.

REFERENCES

1. Kumar A, et al. 2002. *Genes Dev* **16**:707–719.
2. Gindullis F and Meier I. 1999. *Plant Cell* **11**:1117–1128.
3. Gindullis F, et al. 1999. *Plant Cell* **11**:1755–1768.
4. Pandey A and Mann M. 2000. *Nature* **405**:837–846.
5. Westbrook JA, et al. 2006. *Electrophoresis* **27**:1547–1555.
6. Dreger M. 2003. *Eur J Biochem* **270**:589–599.
7. Storrie B and Madden EA. 1990. *Methods Enzymol* **182**:203–235.
8. Park OK. 2004. *J Biochem Mol Biol* **37**:133–138.
9. Fisher PA, et al. 1982. *J Cell Biol* **92**:674–686.
10. Gerner C and Sauermann GJ. 1999. *J Cell Biochem* **71**:363–374.
11. Gerner C, et al. 1999. *J Cell Biochem* **74**:145–151.
12. Mattern KA, et al. 1997. *J Cell Biochem* **65**:42–52.
13. Dreger M, et al. 2001. *Proc Natl Acad Sci USA* **98**:11943–11948.
14. Andersen JS, et al. 2002. *Curr Biol* **12**:1–11.
15. Jung E, et al. 2000. *Electrophoresis* **21**:3483–3487.
16. Bae MS, et al. 2003. *Plant J* **36**:652–663.
17. Khan MK and Komatsu S. 2004. *Phytochemistry* **65**:1671–1681.
18. Calikowski TT, et al. 2003. *J Cell Biochem* **90**:361–378.
19. Pendle AF, et al. 2005. *Mol Biol Cell* **16**:260–269.
20. Pandey A, et al. 2006. *J Prot Res* **5**:3301–3311.
21. Baker DG, et al. 1990. *Plant Mol Biol Rep* **8**:40–49.
22. Anderson LE, et al. 1995. *Plant Physiol* **108**:659–667.
23. Razem FA, et al. 2006. *Nature* **439**:290–294.
24. Brunet A, et al. 2002. *J Cell Biol* **156**:817–828.
25. Kunzler M, et al. 2000. *Mol Cell Biol* **20**:4295–4308.
26. Millar AH, et al. 2001. *Plant Physiol* **127**:1711–1727.
27. Singh R and Green MR. 1993. *Science* **259**:365–368.
28. Nagy M, et al. 1982. *J Cell Biol* **92**:790–794.
29. Grzanka A, et al. 2004. *Oncol Rep* **11**:765–770.
30. Hu P, et al. 2004. *Genes Dev* **18**:3010–3015.

31. Percipalle P, et al. 2003. *Proc Natl Acad Sci USA* **100**:6475–6480.
32. Percipalle P and Visa N. 2006. *J Cell Biol* **172**:967–971.
33. Gindullis F and Meier I. 1999. *Plant Cell* **11**:1117–1128.
34. Gangwani L, et al. 1998. *J Cell Biol* **143**:1471–1484.
35. Nollen EA, et al. 2001. *Proc Natl Acad Sci USA* **98**:12038–12043.
36. Raikhel N and Coruzzi G. 2003. *Plant Physiol* **132**:404–409.

CHLOROPLAST

Thomas Kieselbach and Wolfgang P. Schröder

23.1 INTRODUCTION

Chloroplasts are semiautonomous organelles that are considered to originate from a cyanobacterial ancestor. Their principal task is oxygenic photosynthesis and production of carbohydrates. In addition, their metabolic functions include biosynthesis of lipids, amino acids, chlorophylls, and carotenoids. Figure 23.1 shows that chloroplasts comprise six different compartments. The entire organelle is enclosed in a double envelope that surrounds the stroma, and embedded in the stroma is the thylakoid membrane that in turn surrounds a narrow compartment known as the lumen. The chloroplast genome contains about 100 genes, but the majority of chloroplast proteins are nuclear encoded and, thus, need to be imported from the cytoplasm.

Chloroplasts have been early in the focus of plant proteomics. The first proteome studies of chloroplasts came in time with the completion of the *Arabidopsis* genome project and were important for the development of the field of plant proteomics. The complex ultrastructure of chloroplasts and the broad dynamic range of their proteins make a complete proteome analysis a difficult task. The presence of high abundance proteins, such as the stromal enzyme RuBisCO and the thylakoid-located LHC II,

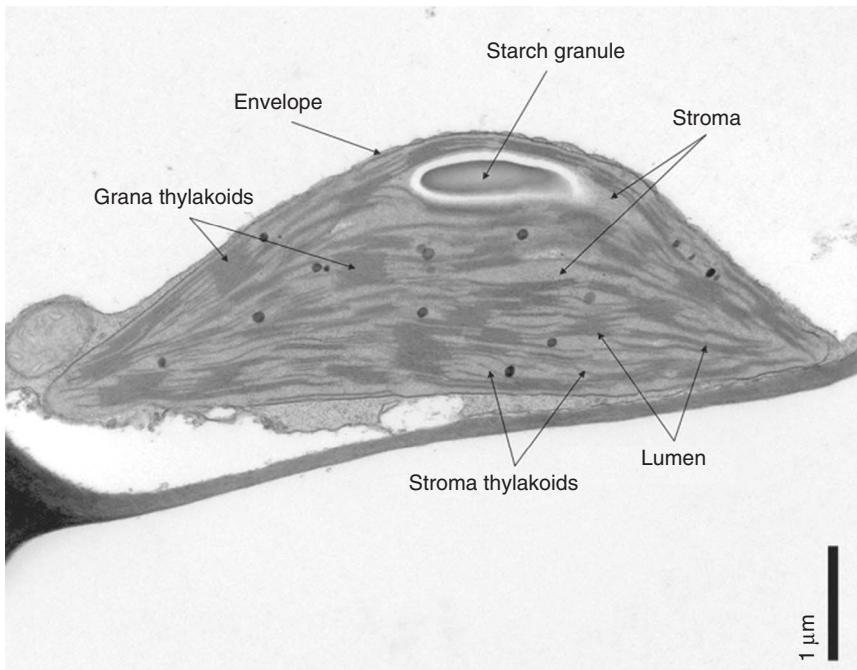


FIGURE 23.1. Electron micrograph showing the ultrastructure of an *Arabidopsis* chloroplast.

makes a comprehensive analysis of low-abundance proteins difficult. These proteins can only be detected, if they are enriched in a specific chloroplast fraction. For this reason, most studies of the chloroplast proteome have focused on the analysis of individual compartments or employed systematic fractionation instead of using the entire organelle.

The sequencing of the *Arabidopsis* genome and the development of bioinformatics tools for a genome-wide analysis of protein localization allowed an estimation of the number of chloroplast proteins. Bioinformatics analyses were also quite frequently included in studies of the chloroplast proteome. The *in silico* work was used as a quality control of the experimental data sets, and it offered a chance to go beyond the limits of experimentation by a prediction of the unidentified proteins of the entire organelle and its individual compartments [1–3]. According to a recent improved prediction, the *Arabidopsis* chloroplast has 2090 proteins [4]. However, the number of experimentally identified chloroplast proteins of *Arabidopsis* is still much lower. Currently, the plastid protein database plprot [5] contains 690 chloroplast proteins of *Arabidopsis* that have been experimentally verified in proteomics studies, and the *Arabidopsis* SUBA [6] has 1110 entries for proteins that were found by MS in plastids. Hence, if the bioinformatics prediction of chloroplast located *Arabidopsis* proteins was approximately correct, only about half of these proteins have been experimentally identified so far. As for rice, the number of predicted chloroplast proteins is 4853 [4]; many of these are considered to be species-specific, which implies that chloroplasts

of different plants might contain different pathways. This situation reflects some of the difficulties of chloroplast proteome analysis, and it shows that there is still a long way to go to understand the proteome of higher-plant chloroplasts.

23.2 BRIEF BIBLIOGRAPHIC REVIEW

A brief glance at the literature in chloroplast proteomics tells us that many studies approached the aim to contribute to a comprehensive catalogue of chloroplast proteins and their subcellular locations. Finding unknown biological functions by an in-depth analysis of individual chloroplast compartments was a major question behind this kind of research. Important contributions include investigations into the luminal chloroplast proteome and peripheral thylakoid proteins [2, 3], as well as detailed proteome analysis of the thylakoid membrane [7, 8] and of the protein complexes of the chloroplast stroma [9]. Proteomics of the envelope provided novel insights into the transporters and the metabolism of this membrane [10–12], and large-scale studies of the *Arabidopsis* chloroplast proteome added to the knowledge of its metabolic pathways and the catalogue of known chloroplast proteins [1, 13, 14]. Moreover, recent proteome studies indicated a metabolic activity of the chloroplast plastoglobules [15, 16]. Because *A. thaliana* provided the benefit of an annotated genome that enabled protein identification by conventional MS, it was the model organism of choice. Hence, chloroplast proteomics has mainly been a study of the *Arabidopsis* chloroplast. Proteomics techniques were also applied to study the subunit composition of photosynthetic protein complexes and to find new phosphorylation sites of thylakoid membrane proteins [17–19]. Combination of proteomics techniques with ligand affinity chromatography proteomics provided important novel insights into the network of thioredoxin-mediated signal transduction [20–23]. More recently, quantitative proteomics contributed to a better understanding of the metabolic differentiation of bundle sheath and mesophyll chloroplasts [24]. In addition, quantitative proteome studies gave insights into key events during early chloroplast biogenesis [25], the response of the chloroplast to low temperature, and the effect of strong light under ascorbate deficiency [26, 27].

23.3 METHODOLOGY AND STRATEGY

One of the central points in any proteomics study is the preparation of the biological samples that are analyzed. Although current biomolecular MS enables very sophisticated applications such as detection of low-abundance proteins in complex mixtures, there are limits for the sensitivity of detection and the dynamic range that need to be considered in the design of a proteomics study. The results of the MS analysis in the end of a work flow cannot be better than the samples that are applied for it. For this reason we focus on the preparation of samples in different chloroplast proteome studies. We also add a brief discussion of the analytical approaches to identify the proteins of the different chloroplast fractions. However, details of conventional proteomics techniques are presented elsewhere in this book.

The experimental design of most proteomics studies includes three steps that contain (i) a sample preparation and (ii) 1D or 2D separation by electrophoresis or chromatography, followed by (iii) MS analysis of the separated proteins. While the last two steps usually involve standard techniques, the first step does not. Any sample preparation starts with the choice of the plant material that is critical for the outcome of a proteome study and its reproducibility. The majority of studies used soil-grown *Arabidopsis* plants. However, plants grown hydroponically or on plates were also used. The age of the plants varied from 3 to 4 up to 7 weeks. With respect to light and illumination times, the variations are somewhat smaller. Normally the light intensity was 100–200 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ with 150 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ in most cases; illumination times were between 8 and 14 h. So far no studies have been conducted on how these parameters affect the chloroplast proteome. The broad range of growth conditions used in different studies suggests that the proteome profiles display snapshots that show parts of the chloroplast proteome at different times of plant development. Figure 23.2 gives an overview of principal steps that were used in the sample preparation of some important studies of the chloroplast proteome. In many studies, Percoll gradient centrifugation was used to purify intact chloroplasts from broken chloroplasts and membranes from other organelles [1, 10, 11, 13–15]. The study of the luminal chloroplast proteome used chloroplasts that were purified by differential centrifugation to prepare a high amount of start material for a further purification of the luminal proteins [2]. To enable identification of low-abundance proteins, the sample preparation included usually the purification of a subcellular

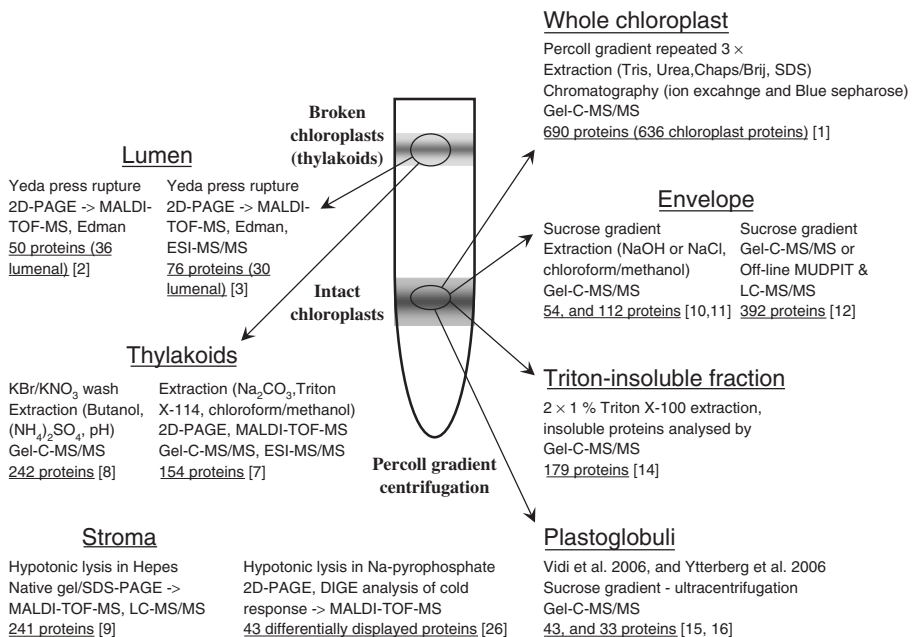


FIGURE 23.2. General methods used in chloroplast proteome studies.

compartment or a systematic fractionation of the organelle. For instance, envelope membranes and plastoglobules were purified using sucrose density centrifugation [10, 11, 15, 16]. Stromal chloroplast fractions were purified by classical hypotonic lysis [9, 26]. Thylakoid membranes were purified by different wash treatments [7, 8]. Rupture of carefully washed thylakoids using a Yeda press was an easy and efficient method to isolate a highly pure fraction of soluble lumen proteins [2]. A similar approach using differently purified thylakoids yielded a fraction of peripheral thylakoid proteins that contained a higher amount of stroma proteins [3].

Because the chloroplast is a membrane-rich organelle (Figure 23.1), specialized techniques were needed to extract the hydrophobic membrane proteins of the envelope and the thylakoid membrane. For instance, C/M solutions have strong interactions with lipids and were therefore useful to extract proteins from the hydrophobic core of the envelope [10]. In combination with alkaline and salt treatments, this method enabled identification of a broad range of the peripheral and integral envelope proteins [11]. A combination of salt and organic solvent treatments resulted also in an effective fractionation of thylakoid membrane proteins [7]. An alternative approach to fractionate the proteins of the thylakoid membrane involved a wash with chaotropic agents (KBR, KNO_3) to remove a large part of the peripheral proteins. Following extraction of the pigments, the remaining peripheral and integral thylakoid proteins were then fractionated using sequential $(\text{NH}_4)_2\text{SO}_4$ precipitation and pH shifts. This approach enabled identification of many known or potential integral thylakoid proteins [8]. The variety of different methods that have been used to extract proteins of different solubility from the envelope, and the thylakoid membrane showed that none of the extractions was complete. Instead a combination of complementary extraction steps had to be performed to cover a broad range of envelope and thylakoid proteins. As for the entire chloroplast, sequential extraction using four different buffer systems including Tris, urea, thiourea, and detergents provided suitable samples for a large-scale identification of chloroplast proteins [1]. An alternative approach removed the major part of the envelope and thylakoid membranes by solubilization with Triton X-100. The fraction of Triton X-100 insoluble proteins comprised a complementary set of chloroplast proteins, many of which had not been identified in the large-scale analysis of the entire chloroplast proteome [14].

An important part in the workflow of proteomics involves separation of the proteins by electrophoresis or chromatography before their analysis by MS. A classical approach combines 2-DGE and MALDI-TOF-MS or ESI-MS. However, this approach was only useful for studies of soluble proteins such as luminal and stromal chloroplast proteomes or peripheral thylakoid proteins [2, 3, 8, 26]. As for membrane proteins, separation by BN-PAGE combined with SDS-PAGE was a successful approach to study assembly and turnover of PS II under light and iron stress conditions [17, 18]. Alternatively membrane proteins were separated by RP chromatography before enzymatic cleavage of the collected protein fractions and their analysis by MS/MS [7, 8]. A common design to analyze membrane proteins involved a 1D separation by SDS-PAGE that was followed by enzymatic or chemical cleavage of the excised protein bands. The resulting peptides were further separated by nRP-HPLC and analyzed on-line by MS/MS (Gel-C-MS/MS). This approach was, for instance,

applied in studies of the chloroplast envelope proteome [10–12], and it enabled the identification of more than 100 proteins in the different fractions of extracted envelope proteins of *Arabidopsis* [11, 12]. A complementary approach for the analysis of envelope proteins involved MudPIT. Instead of a separation of intact proteins, the purified envelope proteins were digested using trypsin and the peptides separated by ion exchange chromatography. The collected peptide fractions were further separated by on-line RP-HPLC, and the eluted peptides were analyzed on-line by MS/MS. The combination of MudPIT with Gel-C-MS/MS enabled the identification of 392 proteins, 149 of which were only found by MudPIT [12]. A series of multidimensional separation steps was also used in the study of the entire chloroplast proteome. Separation of the extracted chloroplast proteins by ion exchange chromatography or Blue Sepharose affinity chromatography was combined with SDS-PAGE to resolve the complex mixtures of chloroplast proteins. Individual fractions and protein bands were cleaved by enzymatic digestion, and the resulting peptides were analyzed by RP-LC-MS/MS. This design enabled the identification of 690 chloroplast *Arabidopsis* proteins, 636 of which were considered to be part of the chloroplast proteome [1].

23.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Chloroplasts have been in the focus of intense research for a long time, and the knowledge of their functions was already quite profound 20 years ago. The biochemical map of chloroplasts displayed in prominent positions the light reactions of photosynthesis, flanked by the Calvin cycle and the pathways for the synthesis of starch and sucrose. Photorespiration and the transport of metabolites across the envelope had been quite thoroughly studied, and much was known about the pathways for the biosynthesis of lipids, amino acids, chlorophylls, and carotenoids. In summary, chloroplasts were a well-studied biological system, when chloroplast proteomics started to develop. That leads to the question: has proteomics contributed to our knowledge of the chloroplast?

An important part of research in chloroplast proteomics aimed to catalogue the composition of the chloroplast proteome and the subcellular locations of its proteins. As presented in the previous section, the studies that contributed to this work either analyzed purified subcellular fractions such as the envelope or the lumen or employed sequential fractionation protocols to overcome the dynamic range problem of the chloroplast. Facing the cytoplasm, the envelope provides the interface to the plant cell, and a major reason for the study of its proteome was to get new insights into its transporters. Considering that the envelope only accounts for a small part of the total protein content of chloroplasts, purification and analysis of its proteins are no easy task. However, three recent studies achieved identification of more than 300 nonredundant proteins in envelope preparations of spinach and *Arabidopsis* [10–12]. As expected, many proteins that were found had putative or known functions as transporters. An important result was, for instance, the identification of novel potential phosphate transporters [11]. All known subunits of the Toc and Tic complexes were identified, and new putative components of the protein import machinery were found. Moreover,

the presence of many unknown proteins indicated that the chloroplast envelope also fulfills functions that have not been discovered yet [10–12]. The assignment of the subcellular location of the novel proteins that were found in the envelope preparations turned out to be complicated. While proteins of the inner envelope usually have a chloroplast targeting peptide and known or putative TM helices, proteins of the outer envelope often do not have a targeting peptide and lack a distinct common marker that can be used for reliable bioinformatics predictions. As a consequence, it was difficult to distinguish putative outer-membrane proteins from cross-contamination [10, 11]. This problem became very clear in one study, because only 113 of 392 identified proteins contained a predicted or known TMD. Hence, the majority of these proteins could only be considered as potential candidate envelope proteins, and their subcellular location needs confirmation by additional experiments [12].

Enclosed in the envelope are the chloroplast stroma that, in turn, encloses the thylakoid membrane and the plastoglobules. The stroma accounts for the major protein content of the chloroplast. However, the high amount of RuBisCO complicates an analysis of its low-abundance proteins, and few studies have addressed the stromal proteome. Using techniques that were developed for the analysis of protein complexes, a recent study achieved identification of more than 200 stromal proteins, most of which had functions in known metabolic pathways. Noteworthy, the majority of these proteins formed homomeric complexes of a single mass, and for about 60% of the identified stroma proteins, assembly into homomeric complexes had also been observed in other studies. However, further investigation is needed to find out if assembly into homomeric complexes, indeed, is a characteristic of the protein interaction network in chloroplasts [9]. As for the plastoglobuli, recent proteomics studies achieved the identification of novel plastoglobuli proteins that included many fibrillins and α -tocopherol cyclase VTE1. While plastoglobules were considered to function mainly in the storage of lipids and quinones, these findings indicated that they also were metabolically active and participated in the synthesis of tocopherol [15, 16]. Embedded in the stroma, the thylakoid membrane is the center of oxygenic photosynthesis and has been studied in great detail. Hence, it provided a challenge for proteomics to add new insights to the profound knowledge of thylakoid functions. Using series of sequential extractions and ammonium sulfate precipitations to fractionate the hydrophobic thylakoid proteins, two complementary studies succeeded in the identification of more than 300 proteins. In addition to many proteins with known functions in light harvesting and photosynthetic electron flow, a large number of proteins with unknown functions were found. Many of these proteins had one or more predicted TM helices, which indicated that they were novel integral thylakoid proteins [7, 8]. In summary, the results of these studies provided evidence that the thylakoid membrane probably had functions that have not been discovered yet.

The thylakoid lumen is the most distant of all chloroplast compartments, and it was an important achievement to discover that this compartment was no empty space but contained many proteins. Further development of this work produced detailed insights into the soluble proteome of the chloroplast lumen. Relevant results were the identification of a specific luminal chloroplast proteome that comprised a relatively small number of protein families. The small family of known luminal chloroplast

proteins grew up to 36 members, and a restrictive bioinformatics analysis of the *Arabidopsis* genome predicted that the chloroplast lumen contained 80 proteins. The distinct characteristics of the luminal proteome included two large families of luminal immunophilins and PsbP domain proteins. Moreover, systematic analysis of the luminal transit peptides showed that 19 of 35 luminal proteins with a known gene model had a twin-arginine motif, showing that the twin-arginine translocation (TAT) system was a principal translocation pathway for protein import into the thylakoid membrane [2]. The luminal chloroplast proteome was also addressed in a study of peripheral thylakoid proteins [3]. In total, 81 proteins were identified that included 30 luminal proteins, of which 24 proteins overlapped with those that were found in the proteome analysis of highly purified chloroplast lumen [2]. An important complement to the experimental work was a genome-wide *in silico* prediction of luminal chloroplast proteins that yielded a list of 71 *Arabidopsis* proteins that had a twin-arginine motif for import by the TAT pathway [3].

The first study of the whole chloroplast proteome of *Arabidopsis* accomplished a successful identification of 690 proteins and provided some surprising results. A set of 636 proteins was considered to be *bona fide* chloroplast proteins, of which 141 proteins had no known functions. The study achieved an almost complete coverage of the enzymes of the Calvin cycle, and it covered 65% of nuclear-encoded photosynthetic electron transport proteins. However, coverage of the less-abundant proteins of the amino acid biosynthesis pathways was only 23%. A complementary profiling of RNA expression revealed a positive correlation of transcript levels and relative protein abundance for some pathways, such as the Calvin cycle, but not for others, which indicated different control mechanisms for different metabolic pathways [1]. This approach was also useful to uncover technical biases in the experimental design and to assess the subcellular location of individual chloroplast proteins [13]. An *in silico* analysis of the identified chloroplast protein using TargetP produced the surprising result that only 376 (59%) of the 636 identified chloroplast proteins were predicted to have a targeting peptide for import into the chloroplast. Considering that 32 proteins were plastid-encoded, 228 proteins remained that had no predicted targeting peptides, which was a much larger number than expected [1]. That raised the question as to whether there was an alternative route for protein import into the chloroplast through the endomembrane system [28].

According to the SUBA database [6] all these studies identified together 919 proteins, 718 of which had an assigned subcellular location in plastids. A comparison with a recent *in silico* prediction of nuclear-encoded chloroplast proteins [4] showed that 425 (59%) of these proteins were present in both the predicted and the experimental protein list. A similar result was obtained using the set of experimental chloroplast proteins of *Arabidopsis* from the plastid proteome database PPDB [7]. If splice forms were omitted, this list contained 676 nuclear encoded, plastid located *Arabidopsis* proteins, of which 474 (70%) were present on the list of the *in silico* predicted chloroplast proteins [4]. In summary, only 60–70% of the experimentally identified chloroplast proteins had a predicted location in plastids. This observation did not come unexpected, because it had turned out that the sensitivity of bioinformatics tools like TargetP to predict chloroplast targeting peptides was lower than originally

reported [4]. Consistent observations were also made in a comparison of predicted and identified proteins in studies of the entire chloroplast proteome [1, 13].

PTMs play an important role for the regulation of protein function, and studies of the chloroplast lumen and the thylakoid membrane showed distinct isoelectric heterogeneity for many luminal proteins such as PsbO and peripheral proteins such as CF1 α,β [2, 7]. However, the type and functions of these modifications still remain to be investigated. One of the best-studied PTMs is phosphorylation, and its impact for the regulation of photosynthesis is well known. A recent refined approach to enrich phosphorylated peptides succeeded in the detection of three new *in vivo* phosphorylation sites in the thylakoid membrane of *Arabidopsis*. Consistent with the knowledge of thylakoid protein phosphorylation, all phosphorylation sites were found in threonine residues. One of the novel phosphoproteins was PsbD, showing for the first time phosphorylation of PS I [19]. Phosphorylation is closely linked to the redox state of the chloroplast, and combination of proteomics techniques with affinity chromatography on modified thioredoxin ligands enabled identification of many new potential thioredoxin targets in the chloroplast of spinach and *Arabidopsis* [20, 23]. There is now convincing evidence that thioredoxin interactions regulate proteins not only in the chloroplast stroma but also in the thylakoid lumen and probably are a feature of the entire chloroplast [21–23]. Among the novel thioredoxin targets are proteins with central functions such as PsbO of PS II²² and violaxanthin de-epoxidase of xanthophylls (Kieselbach, unpublished results).

Quantitative techniques gave chloroplast proteomics a new dimension by providing tools to compare proteomes of different chloroplast types and different time points in dynamics of chloroplast proteomes. For instance, a quantification of metabolic pathways in bundle sheath and mesophyll chloroplasts of maize showed that their metabolic differentiation was consistent with their different roles in C4 photosynthesis. Mesophyll chloroplasts accommodated primarily pathways that need NADPH such as, for instance, nitrogen import and biosynthesis of lipids. By contrast, proteins that participate in the Calvin cycle and the synthesis of starch were primarily accumulated in bundle sheath chloroplasts [24]. In addition to a static comparison of different proteomes, quantitative methods can provide insights into the dynamic response of chloroplasts to changing environmental conditions. A study of cold response of *Arabidopsis* chloroplasts showed that short-term acclimation (10 days) resulted in major changes in the stromal proteome, while long-term acclimation (40 days) induced modulation of the proteomes of both the chloroplast lumen and the stroma [26]. A comparative study of the *Arabidopsis* mutant *vct2-2* and wild-type plants addressed the impact of ascorbate deficiency on the dynamics of the chloroplast proteome under high light stress. Ascorbate levels in *vct2-2* leaves were about 20% of wild-type levels. During five days of illumination with 1000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ *vct2-2* and wild-type plant, significant differences in the relative abundance of FeSOD, Cu/Zn-SOD, Hsp70, PsbS, and chloroplast glyoxylate were displayed [27]. Quantitative proteomics provided also insights into the light-induced greening of etioplasts. This process involves major changes in the plastid proteome to switch from the heterotrophic metabolism of etioplasts to the photoautotrophic metabolism of chloroplasts. A recent study in rice showed that the early phase of etioplasts differentiation was marked by a distinct

accumulation of Calvin cycle enzymes and thylakoid membrane proteins that were needed for the construction of the photosynthetic apparatus. The strongest increase was observed for RuBisCO. The dynamics of the etioplasts proteome displayed also significant up-regulation of proteins that controlled the translation of mRNA. In addition, light-induced phosphorylation of a plastid ribonucleoprotein (RNP) might stabilize mRNA that is needed for protein synthesis [25].

23.5 CONCLUSIONS

During recent years, chloroplast proteomics has become an established research discipline. As for *Arabidopsis*, more than 700 chloroplast proteins have been identified that provide insights into the proteomes of all chloroplast compartments. In addition to cataloguing chloroplast proteins, quantitative studies provided first insights into proteome dynamics during plastid differentiation and response to changing environmental conditions. The results of individual studies have demonstrated that the chloroplast proteome cannot be analyzed in a single approach but that many complementary techniques are required to cope with the broad dynamic range of chloroplast proteins and the special properties of the envelope and thylakoid membrane. While the results of the proteome analysis of *Arabidopsis* chloroplasts include many proteins with known and expected functions, a considerable part of the novel chloroplast proteins have no known functions. That implies not only that the *Arabidopsis* chloroplast has unknown functions, but also that many details about the regulation of known pathways remain to be investigated. Providing a large resource of candidate proteins for functional studies has been one of the principal results of chloroplast proteomics. The studies of the chloroplast proteome have also clearly shown the current limitations of bioinformatics predictions of plastid proteins. Only 60–70% of the identified chloroplast proteins of *Arabidopsis* had also a predicted location in plastids. In summary, chloroplast proteomics has so far not uncovered any novel principal functions of plastids but has contributed a valuable resource of knowledge for further functional studies.

23.6 FIVE-YEAR VIEWPOINT

Chloroplast proteomics has so far mainly focused on *A. thaliana*, and there is still little known about the chloroplast proteome of other plants. A comparison of rice and *Arabidopsis* could provide important insights, if chloroplasts from different species have differences in their functions. Improved techniques for protein and peptide separation in large-scale studies and software for database-independent interpretation of MS/MS spectra will probably enable a more complete analysis of the chloroplast proteome of *Arabidopsis* and other model plants. The currently known chloroplast proteins provide also a useful resource for an improvement of bioinformatics prediction tools, and the confidence of plastid protein prediction might become as high as 90% or better. In addition, the need for revisiting the genome annotation of *Arabidopsis* has become more and more clear, and a new version of the *Arabidopsis* genome with improved

gene models might be released, which would not only increase proteome coverage of experimental studies but would also provide a better base for bioinformatics predictions. In addition, the dynamics of the chloroplast proteome during different stages of chloroplast biogenesis and under changing environmental conditions, such as temperature and light, will probably move more into the focus of quantitative proteomics and contribute to a deeper knowledge of plastid functions.

REFERENCES

1. Kleffmann T, et al. 2004. *Curr Biol* **14**:354–362.
2. Schubert M, et al. 2002. *J Biol Chem* **277**:8354–8365.
3. Peltier JB, et al. 2002. *Plant Cell* **14**:211–236.
4. Richly E and Leister D. 2004. *Gene* **329**:11–16.
5. Kleffmann T, et al. 2006. *Plant Cell Physiol* **47**:432–436.
6. Heazlewood JL, et al. 2007. *Nucleic Acids Res* **35**:D213–D218.
7. Friso G, et al. 2004. *Plant Cell* **16**:478–499.
8. Peltier JB, et al. 2004. *J Biol Chem* **279**:49367–49383.
9. Peltier JB, et al. 2006. *Mol Cell Proteomics* **5**:114–133.
10. Ferro M, et al. 2002. *Proc Natl Acad Sci USA* **99**:11487–11492.
11. Ferro M, et al. 2003. *Mol Cell Proteomics* **2**:325–345.
12. Froehlich JE, et al. 2003. *J Proteome Res* **2**:413–425.
13. Baginsky S, et al. 2005. *J Proteome Res* **4**:637–640.
14. Phinney BS and Thelen JJ. 2005. *J Proteome Res* **4**:497–506.
15. Vidi PA, et al. 2006. *J Biol Chem* **281**:11225–11234.
16. Ytterberg AJ, Peltier JB, and van Wijk KJ. 2006. *Plant Physiol* **140**:984–997.
17. Aro EM, et al. 2005. *J Exp Bot* **56**:347–356.
18. Andaluz S, et al. 2006. *Photosynth Res* **89**:141–155.
19. Hansson M and Vener AV. 2003. *Mol Cell Proteomics* **2**:550–559.
20. Balmer Y, et al. 2003. *Proc Natl Acad Sci USA* **100**:370–375.
21. Balmer Y, et al. 2006. *Antioxid Redox Signal* **8**:1829–1834.
22. Marchant C, Le Marechal P and Meyer Y. 2004. *Proteomics* **4**:2696–2706.
23. Marchand C, et al. 2006. *Proteomics* **6**:6528–6537.
24. Majeran W, et al. 2005. *Plant Cell* **17**:3111–3140.
25. Kleffmann T, et al. 2007. *Plant Physiol* **143**:912–923.
26. Goulas E, et al. 2006. *Plant J* **47**:720–734.
27. Giocamelli L, Rudella A and van Wijk KJ. 2006. *Plant Physiol* **141**:685–701.
28. Jarvis P. 2004. *Curr Biol* **14**:R317–R319.

ETIOPLAST

Anne von Zychlinski, Sonja Reiland, Wilhelm Gruissem,
and Sacha Baginsky

24.1 INTRODUCTION

Plastids are plant cell organelles that have important biosynthetic and metabolic functions, which include photosynthetic carbon fixation and the synthesis of amino acids, fatty acids, starch, and secondary metabolites such as pigments [1, 2]. As reflected in the diverse set of plastid metabolic activities, plastids are a heterogeneous group of cell organelles that can develop and differentiate from undifferentiated precursor plastids (proplastids) into specialized plastid types. Proplastids in leaf tissues can develop either into photosynthetically active chloroplasts or into nonphotosynthetic etioplasts, depending on the availability of light. Etioplasts, which develop in the absence of light, are nonphotosynthetic plastids that contain a large paracrystalline structure, which consists of protochlorophyllide oxidoreductase (POR), NADPH, and protochlorophyllide. Upon illumination of dark-grown plants, POR is activated and catalyzes the reduction of protochlorophyllide to chlorophyllide, which is subsequently converted into chlorophyll. Within a few hours of illumination, the etioplast prolamellar body is fully disassembled and converted into a functional thylakoid membrane

system responsible for photosynthetic electron transport. This differentiation system is tuned for a rapid response to light because chloroplast formation and the energy from photosynthesis are important to sustain cellular functions of the plant cell.

Light-induced etioplast-to-chloroplast conversion is a complex process that involves the reorganization of the etioplast proteome to support photosynthesis-dependent autotrophy. Some aspects of such a transition have been investigated, but to date these were mostly restricted to the induction of individual proteins and the synthesis of chlorophyll and other pigments [reviewed in reference 2]. A more comprehensive analysis of this transition provides deeper insights into the reorganization of the proteome and the changes in metabolite fluxes and regulatory cross-talk between different metabolic functions during plastid differentiation. It is conceivable that the global rearrangement of a heterotrophic metabolism to autotrophy must involve most of the essential metabolic functions to ascertain a balanced shuffling of energy and reducing equivalents between different metabolic activities that constitute the plastid metabolism. In this chapter, we present the results of recent comprehensive proteome analyses of rice etioplasts and provide insights into the initial reorganization of the etioplast proteome during etioplast-to-chloroplast conversion on the basis of quantitative proteomics data.

24.2 ETIOPLAST PROTEOME ANALYSIS

A Novel Protocol for Isolation of Intact and Pure Etioplasts from Rice Leaf Blades

Proteome analyses of isolated cell organelles provide new information about protein localization, intracellular sorting, and pathway compartmentalization. In order to acquire reliable results about the latter, it is absolutely necessary that the organelles used for the proteome analysis are free from contaminations with other cell organelles, or that at least the degree of contamination can be assessed quantitatively. We therefore devised a new strategy for the isolation of etioplasts, which is based on the combination of several differential centrifugation steps with Nycodenz density gradient centrifugation and assessed the degree of potential contamination by semiquantitative protein profiling along the density gradient (Figure 24.1). Characteristic marker proteins for each cell organelle suggested that cytosolic proteins, vacuolar proteins, and proteins from the ER together with some mitochondrial and nuclear proteins migrate at the top of the gradient (Figure 24.1; 10% Nycodenz). Peroxisomal proteins, some cytosolic proteins, and PM fragments migrate at the interphase between 10% and 15% Nycodenz; intact mitochondria and nuclei migrate to the interphase between 25% and 30% Nycodenz; and other PM fragments, some nuclei, peroxisomes, and glyoxisomes are found in the sediment. Intact etioplasts migrate to the 15–20% and the 20–25% interphase, where they are free from proteins from other cell organelles [3].

Employed Strategies for the Analysis of Etioplast Proteome

Proteins for proteome analyses were extracted from the isolated etioplasts by solubilization of the etioplasts in detergent containing buffer. In order to increase the

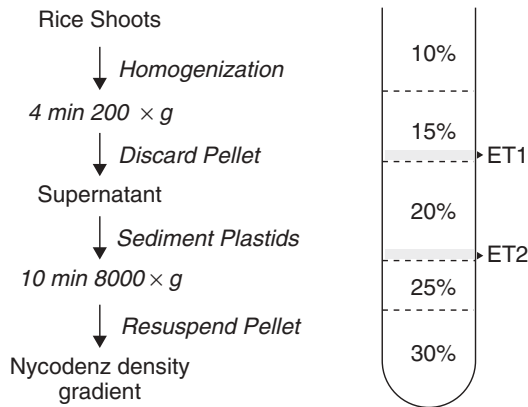


FIGURE 24.1. Purification strategy for rice etioplasts. The purification of rice etioplasts initiated with two differential centrifugation steps followed by Nycodenz density gradient centrifugation. The Nycodenz density gradient consisted of five steps of 10%, 15%, 20%, 25%, and 30% Nycodenz in a standard plastid isolation buffer as described [3]. Fractions at the interphase of each step were collected and subjected to MS protein identification. Etioplast proteins were found in two bands at the interphases of 15% to 20% and 20% to 25% Nycodenz (labeled ET1 and ET2). Both fractions were essentially free from contaminations with proteins of other cell compartments, which were identified in the other fractions of the gradient. We collected etioplasts from the band labeled ET1 and used them for further analysis.

number of detectable etioplast proteins, we employed two different proteomics strategies: shotgun proteomics and 2-DGE. “Shotgun proteomics” attempts the identification of all proteins that are present in a complex sample and results in a list of proteins. Combined with complex protein or peptide fractionation strategies, shotgun proteomics is a technically relatively simple method for high-throughput analyses of the proteome of an organelle or a cell type and provides a snapshot of the major protein constituents [4–6]. Shotgun proteomics has the advantage that it can be combined with 1D SDS-PAGE, making hydrophobic membrane proteins amenable to MS analyses, eliminating protein solubility constraints for proteome analyses. One of the main disadvantages of shotgun proteomics is that it provides only very limited quantitative information about the identified proteins. One possibility to circumvent this shortcoming is spectral or peptide counting because the number of detected peptides per protein relates to the abundance of the protein [reviewed in reference 7]. As an alternative approach, we employed 2-DGE, which provides accurate quantitative information for every identified protein. Furthermore, also PTMs that affect the pI of a protein will be detectable. One disadvantage of 2-DGE is its low resolution of hydrophobic membrane proteins. In conclusion, both strategies have advantages and disadvantages and the combination of both allows increased proteome coverage and a broader scope of the analysis (for further details see Chapters 2 and 3).

The Etioplast Proteome Reveals a Complex Anabolic and Catabolic Metabolism

The results from the two proteome analysis strategies described above were combined and altogether gave rise to 477 nonredundant protein identifications [3, 8]. Figure 24.2 summarizes some of the data on a detailed map of biochemical pathways characteristic for the heterotrophic etioplast. These pathways are important for the plant cell because they provide energy and reducing equivalents for anabolic activities and especially for the synthesis of amino acid. Similar to other heterotrophic plastid types, etioplasts depend on the import of cytosolic sugar phosphates and ATP for their energy metabolism [1, 9, 10]. Our proteomics results are in line with recent biochemical data, which show that etioplasts import triosephosphates via a triosephosphate translocator (TPT) but do not import hexose phosphates [1, 3]. This sugar phosphate import capacity distinguishes etioplasts from heterotrophic BY2 plastids. In the latter, we detected

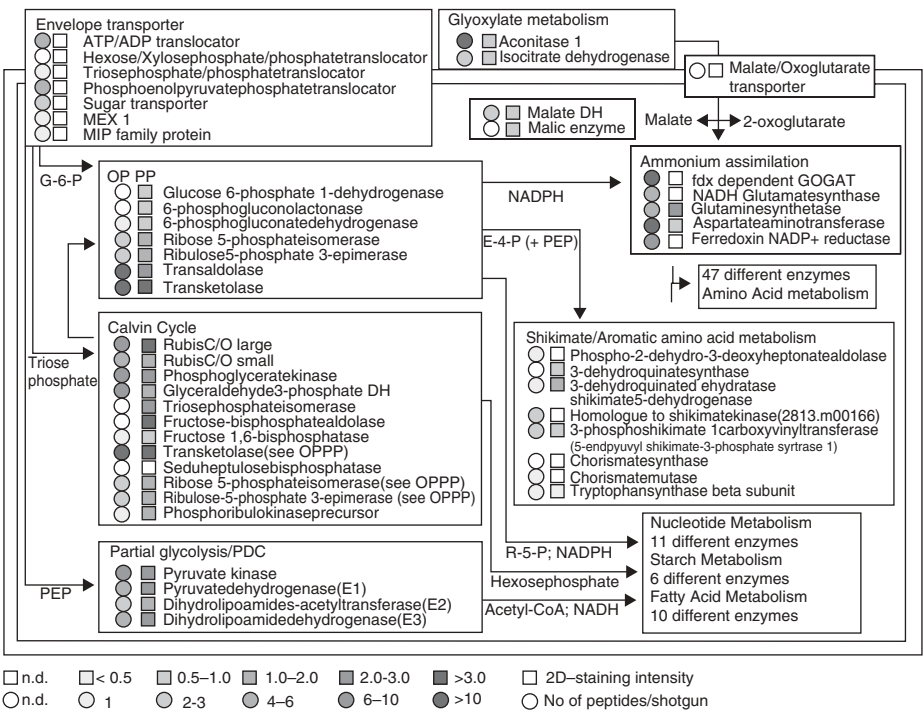


FIGURE 24.2. Color-coded schematic illustration of characteristic metabolic pathways in heterotrophic plastids. Missing protein identifications from known plastid pathways are depicted by white boxes and circles, and abundances of identified proteins are provided as the number of peptides in a shotgun proteome analysis (circles) and the staining intensity in the 2D-map (squares). E-4-P, erythrose-4-phosphate; G-6-P, glucose-6-phosphate; OPPP, oxidative pentose phosphate pathway; PDC, pyruvate dehydrogenase complex; PEP, phosphoenole pyruvate; R-5-P, ribulose-5-phosphate. See insert for color representation of this figure.

high concentrations of the glucose 6-phosphate/phosphate translocator but no TPT [11]. Glucose 6-phosphate enters the oxidative pentose phosphate pathway (OPPP), in which it is metabolized to NADPH and precursors for anabolic pathways—for example, erythrose 4-phosphate for the aromatic amino acid metabolism and ribulose 5-phosphate for the nucleotide metabolism [1, 12]. Because these are important and abundant etioplast metabolic functions (Figure 24.2) and we identified the complete set of enzymes that are active in the OPPP (Figure 24.2), we can reasonably conclude that this pathway must be active.

One of the key regulatory enzymes for the conversion of triosephosphates into hexosephosphates is fructose 1,6-bisphosphatase (FBPase), which is activated by reduced thioredoxin [13, 14]. The discovery of a ferredoxin/thioredoxin redox system in etioplasts and other heterotrophic plastids such as amyloplasts [15, 16], together with the detection of FBPase in etioplasts [3, 8], supports the existence of redox-controlled reactions also in nonphotosynthetic plastids [15–17] (Figure 24.2). This is remarkable, because redox regulation was thought to be restricted to photosynthetic chloroplasts that reduce ferredoxin and thioredoxin via their photosynthetic electron transport. The detection of a complete set of enzymes that enable etioplasts to catalyze the conversion of triosephosphates into hexosephosphates strongly suggests that etioplasts feed glucose 6-phosphate into the OPPP (Figure 24.2). As depicted on the map, the etioplast contains a set of proteins for essential anabolic pathways that provide the plant cell with metabolites. These biosynthetic functions require energy and reducing equivalents that are generated not only by the heterotrophic metabolism in the plastid but also by import of ATP from the cytosol. We identified an ADP/ATP translocator that may play an important role in the shuttling of energy-rich molecules between the cytosol and the etioplast (Figure 24.2).

Ammonia assimilation and the synthesis of amino acids are important plastid functions. We identified all enzymes of the GS/GOGAT (glutamine synthetase/glutamate synthase) cycle in the rice etioplast and, furthermore, several enzymes for the synthesis of other amino acids such as Thr, Lys, Met, His, the branched-chain amino acids, and nearly all enzymes of the shikimic acid pathway (Figure 24.2). Because all proteomics technologies preferentially identify abundant proteins [18, 19], we find only partial coverage of enzymes from other plastid-specific pathways. The identification of several enzymes downstream of chorismate mutase or those active in starch and fatty acid metabolism, however, confirm that other pathways are also present and most likely active in the etioplast but that the concentration of their constituents is rather low.

Plastid Protein Targeting

The plastid protein import machinery is well characterized and most proteins of the translocon complex in the outer (Toc) and inner (Tic) envelope membrane are now functionally characterized [reviewed in references 20 and 21]. In the shotgun approach, we identified the TM proteins Toc159, Tic110, and Tic40 and the soluble component pitrilylsine (stromal processing peptidase). The detection of Tic40 is remarkable because this protein was not detected in a comprehensive proteome analysis of chloroplasts from 7-week-old *Arabidopsis* plants [22], but was identified in chloroplasts from

younger plants [23, 24]. Furthermore, recent studies have shown that Tic40 is not essential for protein translocation [25] and appears to have a modulating rather than a mechanistic function in plastid protein import. The temporal expression of Tic40 and its detection in the etioplast suggest, however, that the protein may be necessary to modulate the transport mechanism under conditions in which the demand for protein import is changing—for example, during the transition of etioplasts to chloroplasts. In addition to the well-established TOC and TIC components, we identified in rice one protein with similarity to the *Arabidopsis* Tim17/Tim22/Tim23, which can be excluded as a contamination [3]. The identification of a Tim17/Tim22/Tim23 family protein in the etioplast envelope membrane, together with the differential identification of Tic40 in different plastids, suggests a previously unrecognized complexity and dynamic composition of the plastid protein import machinery.

In line with recent proteome analyses of chloroplasts and the detection of novel protein import routes into plastids, we also identified a number of proteins from etioplasts that are not predicted to localize to plastids with software-based targeting prediction tools. This FN prediction of true plastid proteins can have several reasons. One reason for the underestimation of plastid proteome complexity can be alternative protein import routes such as those via the secretory pathway [26]. Another reason lies in different transit peptide structures and specificities of import receptors. These could have evolved to ascertain the import of low-abundance proteins in the presence of highly abundant photosynthetic proteins [19, 27, 28]. In order to identify differences in transit peptide composition between identified rice etioplast proteins [3, 8] and *Arabidopsis* chloroplast proteins [22], we aligned the N-terminal targeting sequences (50 amino acids) of rice and those of *Arabidopsis*. The statistical evaluation supports the view that the difference in targeting prediction, at least in part, stems from different transit peptide compositions, because the N-termini from rice are enriched in the unpolar/aliphatic amino acids Ala and Leu whereas those of *Arabidopsis* are enriched in the hydroxylated amino acid Ser (Figure 24.3) [8].

Proteome Dynamics during Light-Induced Chloroplast Development

One of the hallmarks of chloroplast development is the synthesis and accumulation of chlorophyll, which is tightly controlled because of the highly reactive nature of light-activated chlorophyll. Furthermore, intermediates of the chlorophyll biosynthesis pathway such as Mg-protoporphyrin IX and Mg-protoporphyrin 6-monomethyl ester play a fundamental role in the signaling between plastid and nucleus by restraining the expression of nuclear genes for photosynthetic proteins [29, 30]. We therefore analyzed whether the abundance of enzymes active in this pathway could be regulated by illumination. With the exception of glutamyl-tRNA reductase (GluTR), all known enzymes of the tetrapyrrole pathway were detected. This suggests that they accumulate to moderately high levels already in the dark. Although the flux through this pathway is significantly increased by light [31], no significant changes of enzyme abundance were visible during the first 4 h of illumination. Our quantitative proteomics data therefore suggest that the accumulation of the GluTR in the dark is sufficient to support the rapid increase in pathway flux in response to light and that this flux is

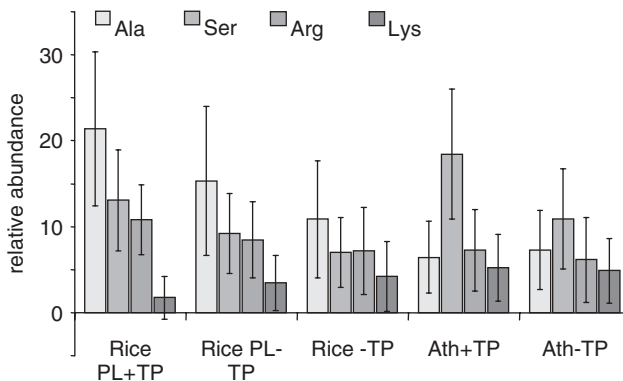


FIGURE 24.3. Statistical analysis of N-termini from rice and *Arabidopsis* proteins with (+ TP) or without a predictable plastid transit peptide (– TP). This figure shows the frequencies in which alanine, serine, arginine, and lysine occur in the 50 N-terminal amino acids of selected rice and *Arabidopsis* proteins. Rice PL + TP, identified rice plastid proteins with a predicted transit peptide; Rice PL – TP: identified rice plastid proteins without a predictable transit peptide; Rice – TP, randomly selected rice proteins without a transit peptide; ATH + TP: randomly selected *Arabidopsis* proteins with predictable transit peptide; ATH – TP: randomly selected *Arabidopsis* proteins without a predictable transit peptide.

likely to be increased by a modification of the enzymes catalytic activity rather than their abundance.

An efficient translation machinery is fundamental for the reorganization of proteomes during developmental processes, and it necessary for the fast accumulation of enzymes that are needed in photosynthetic chloroplasts. Not surprisingly, we detected an increased accumulation of proteins involved in plastid translation after illumination. Among these proteins are several ribosomal proteins and the translation EFs P and Tu. Both EFs have an important function in translation, and it is therefore conceivable that their availability is crucial for rapid and efficient translation in plastids. In addition to an increased accumulation of proteins involved in translation, we find a strong light-induced accumulation of plastid proteases, especially of those that belong to the Clp protease system (Figure 24.4). It is plausible that the degradation of etioplast proteins during chloroplast development fills the plastid amino acids reservoir, which could explain why the relative abundance of amino-acid-synthesizing enzymes decreases in the early illumination phase (Figure 24.4). In contrast to the enzymes of the amino acid metabolism, Calvin cycle enzymes and thylakoid membrane proteins accumulate to higher levels in the light because they are required for establishing the photosynthetic machinery (Figure 24.4).

Light not only regulates the expression of nuclear genes but also affects the plastid gene expression machinery. Light regulation of plastid gene expression occurs at different levels including transcription and post-transcriptional events such as translation and the control of mRNA stability [32–34]. It was previously suggested that plastid mRNA-binding proteins are involved in the regulation of RNA stability

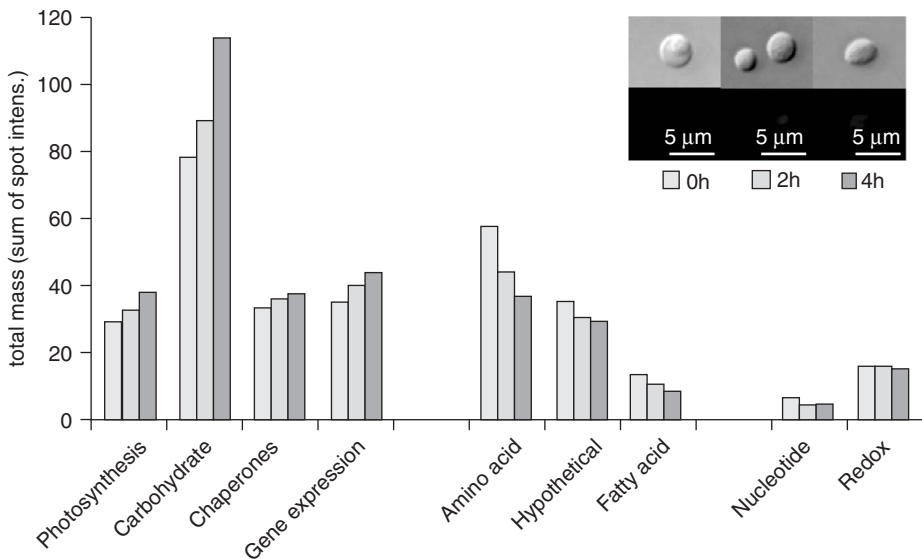


FIGURE 24.4. Reorganization of the etioplast metabolism by illumination. The total protein staining intensities of proteins from each functional category as identified from 2D gels of etioplasts (dark) were summed up and compared with those in each functional category after 2 and 4 h of illumination. The inset shows light microscopy pictures of isolated plastids and their chlorophyll fluorescence 0, 2, and 4 h after the start of the illumination.

[35, 36]. We have identified a number of RNA binding proteins, two of which were found to be phosphorylated. During illumination, the relative amount of phosphorylated RNP increases, suggesting that the illumination induces RNA-binding protein phosphorylation. This finding is in line with the hypothesis that phosphorylation of RNA-binding proteins alters their affinity for RNA and that in this way the stability of plastid mRNAs to external signals is adjusted [34, 37]. In conclusion, our data suggest that adjustments of mRNA stability are among the first events at the gene expression level that initiate the development of chloroplasts from etioplasts.

24.3 CONCLUSIONS

With the development of new protocols for the isolation of intact and pure etioplasts from rice leaf tissue, this plastid type has become amenable for proteome analysis. Using two complementary proteomics strategies, we identified 477 proteins from rice etioplasts, making the etioplast the best-characterized heterotrophic plastid type analyzed to date. The comprehensive analysis of chloroplast development from etioplasts by 2-DGE showed that this transition is associated with a significant increase in proteins for photosynthesis and enzymes involved in the Calvin cycle. At the regulatory level, proteins with a function in translation of plastid mRNAs are up-regulated and mRNAs are stabilized by light-induced phosphorylation of a plastid RNA-binding

protein. These data suggest that plastid translation rates increase upon illumination such allowing for a rapid assembly of a functional thylakoid membrane system. We suggest that most of the energy required during the early phase of chloroplast development is generated by the OPPP and a partial glycolysis module. Further experiments, especially the analysis of metabolite fluxes, will be necessary to expand this model.

24.4 FIVE-YEAR VIEWPOINT

Understanding metabolic pathway compartmentalization and plant cell organelle function are important cornerstones on the way toward plant systems biology (see also Chapter 46). The most successful plant systems biology approaches of the near future will follow a bottom-up strategy ranging from the analysis of individual metabolic modules to an integrative analysis of cell organelle communication. As one of the central nodes in plant metabolism, the plastid will remain a prominent target for systems-level research. The next steps will comprise the acquisition of more quantitative proteomics data in order to understand the regulation of the plastid proteome and the generation of high-quality metabolite profiles in order to assess proteome performance.

REFERENCES

1. Neuhaus HE and Emes MJ. 2000. *Annu Rev Plant Physiol Plant Mol Biol* **51**:111–140.
2. Lopez-Juez E and Pyke KA. 2005. *Int J Dev Biol* **49**:557–577.
3. von Zychlinski A, et al. 2005. *Mol Cell Proteomics* **4**:1072–1084.
4. Washburn MP, et al. 2003. *Proc Natl Acad Sci USA* **100**:3107–3112.
5. Washburn MP, Wolters D, and Yates JR 3rd. 2001. *Nat Biotechnol* **19**:242–247.
6. Koller A, et al. 2002. *Proc Natl Acad Sci USA* **99**:11969–11974.
7. Greenbaum D, et al. 2003. *Genome Biol* **4**:117.
8. Kleffmann T, et al. 2007. *Plant Physiol* **143**:912–923.
9. Weber AP, Schwacke R, and Flugge UI. 2005. *Annu Rev Plant Biol* **56**:133–164.
10. Fischer K and Weber A. 2002. *Trends Plant Sci* **7**:345–351.
11. Baginsky S, Siddique A, and Grissem W. 2004. *J Proteome Res* **3**:1128–1137.
12. Schnarrenberger C, Flechner A, and Martin W. 1995. *Plant Physiol* **108**:609–614.
13. Scheibe R. 1994. *Naturwissenschaften* **81**:443–448.
14. Scheibe R, et al. 2005. *J Exp Bot* **56**:1481–1489.
15. Balmer Y, et al. 2006. *Proc Natl Acad Sci USA* **103**:2988–2993.
16. Balmer Y, et al. 2006. *J Exp Bot* **57**:1591–1602.
17. Schurmann P. 2003. *Antioxid Redox Signal* **5**:69–78.
18. Corbin RW, et al. 2003. *Proc Natl Acad Sci USA* **100**:9232–9237.
19. Baginsky S, et al. 2005. *J Proteome Res* **4**:637–640.
20. Soll J and Schleiff E. 2004. *Nat Rev Mol Cell Biol* **5**:198–208.
21. Kessler F and Schnell DJ. 2006. *Traffic* **7**:248–257.

22. Kleffmann T, et al. 2004. *Curr Biol* **14**:354–362.
23. Ferro M, et al. 2003. *Mol Cell Proteomics* **2**:325–345.
24. Froehlich JE, et al. 2003. *J Proteome Res* **2**:413–425.
25. Ko K, et al. 2005. *J Biol Chem* **280**:215–224.
26. Villarejo A, et al. 2005. *Nat Cell Biol* **7**:1224–1231.
27. Bruce BD. 2001. *Biochim Biophys Acta* **1541**:2–21.
28. Jarvis P and Robinson C. 2004. *Curr Biol* **14**:R1064–R1077.
29. Strand A, et al. 2003. *Nature* **421**:79–83.
30. Surpin M, Larkin RM, and Chory J. 2002. *Plant Cell* **14**:S327–S338.
31. Grimm B. 1998. *Curr Opin Plant Biol* **1**:245–250.
32. Baginsky S and Grissem W. 2002. *Nucleic Acids Res* **30**:4527–4533.
33. Barnes D and Mayfield SP. 2003. *Antioxid Redox Signal* **5**:89–94.
34. Bollenbach TJ, Schuster G and Stern DB. 2004. *Prog Nucleic Acid Res Mol Biol* **78**:305–337.
35. Hayes R, Kudla J, and Grissem W. 1999. *Trends Biochem Sci* **24**:199–202.
36. Schuster G, Lisitsky I, and Klaff P. 1999. *Plant Physiol* **120**:937–944.
37. Lisitsky I and Schuster G. 1995. *Nucleic Acids Res* **23**:2506–2511.

THE PLANT MITOCHONDRIAL PROTEOME AND THE CHALLENGE OF HYDROPHOBIC PROTEIN ANALYSIS

Yew-Foon Tan and A. Harvey Millar

25.1 INTRODUCTION

Mitochondria are ubiquitous organelles that are fundamentally involved in fulfilling the energy requirements of eukaryotic cells. The oxidation of fatty acids, the carbon skeletons of amino acids, and glycolytic end-products through the tricarboxylic acid cycle is linked to the synthesis of NAD(P)H that is subsequently oxidized by the electron transport chain. Concomitant proton translocation is then coupled to the synthesis of ATP in a process known as oxidative phosphorylation. For these processes to occur, the mitochondrion consists of four functionally specialized compartments. The pore-containing outer membrane, surrounding the intermembrane space provides an interactive space between the mitochondrion and the intracellular environment. Glycolytic enzymes have been shown to associate with the outer membrane providing a close linkage of glycolysis with the tricarboxylic acid cycle [1]. The intermembrane space is a domain where mitochondrial-targeted, cytosol-synthesized precursor proteins passage for subsequent integration into the inner membrane or transport into the matrix.

The inner membrane is perhaps the most functionally critical compartment; it has three vital roles in maintaining mitochondrial metabolism. The first is the regulation of communication with the intracellular milieu, because this protein-rich membrane forms an impermeable barrier. The movement of substances across the inner membrane is controlled by a wide variety of carriers and transporters, but many remain poorly characterized to date in plants. The second role is the encapsulation of the matrix where the mitochondrial genome, translational machinery, the tricarboxylic acid cycle, and associated enzymes of carbon and amino acid metabolism are housed. Lastly, the cristae membrane structure of the inner membrane provides a large surface area for biochemical reactions to occur as well as maximizing space for the embedding of respiratory complexes participating in oxidative phosphorylation.

Despite the conserved function and basic structure of the organelle, plant mitochondria have many metabolic discrepancies from their eukaryotic counterparts. The biosynthesis of vital cofactors such as ascorbate, biotin, and folate are notable examples [2–4]. In photosynthetic tissues the carbon metabolism of plant mitochondria is largely given over to its role in the photorespiratory conversion of Gly to Ser [5]. In addition, the presence of an cyanide-insensitive oxidase and additional NADH dehydrogenases provides nonphosphorylating respiratory bypasses that are putatively advantageous as stress adaptations [6, 7]. Plant mitochondria also possess larger genomes than do nonphotosynthetic organisms, which are hypothesized to be relics of their endosymbiotic origins and subsequent evolution.

The theory of mitochondrial origin suggests the progenitor bacteria transferred the majority of its genetic material to the host nucleus over time, rendering the mitochondrion functionally dependent on the host cell [8]. Estimates suggest that plant mitochondria derive 95% of their putative 2000 gene products from the nuclear genome [8, 9]. Hence, a fuller understanding of mitochondrial physiology has been limited by a lack of information regarding the identity of many of the protein constituents of the organelle. Studies into protein targeting lead to the discovery of N-terminal amphiphilic targeting sequences that were often associated with nuclear-encoded, cytosol-synthesized precursor polypeptides destined for mitochondria [10]. This led to the development of numerous protein targeting prediction algorithms to deduced location from primary sequence analysis, but the complexities of gene product localization still requires experimental verification because these programmes have significant FP and FN rates [11]. Transient or constitutive expression of fusion of proteins of interest with fluorescent proteins has been a valuable breakthrough, but as yet no suitable approach for implementation as a high-throughput analysis of the 10,000 s of proteins in plants has been devised.

In recent times, with the availability of annotated genome sequences for several plants and the advances in peptide MS, proteomics has emerged as an invaluable tool in the high-throughput and systematic delineation of the protein constituents of plant mitochondria. In the brief bibliographic review, major aspects of plant mitochondrial proteome research in the post-genomic era will be summarized. The remainder of the chapter will discuss methodologies, results, and applications using proteomics in characterizing the enigmatic hydrophobic proteome of the plant mitochondria.

25.2 BRIEF BIBLIOGRAPHIC REVIEW

The “Global” Mitochondrial Proteome

The initial systematic proteomic analysis of the *Arabidopsis* cell culture mitochondria was independently conducted by two different research groups using the classic 2D IEF/SDS-PAGE approach. CBB-staining of 2D IEF/SDS-PAGE using nonlinear pH gradients of 3–10 reproducibly revealed approximately 350 spots, with 100 expressed in high abundance, while silver-staining of similar gels resulted in the detection of approximately 650 spots [12, 13]. The full number of nonredundant proteins that these spots represented is unknown, because less than 100 nonredundant proteins could be identified by MS or Edman degradation in each study. The classes of proteins identified related to the fundamental mitochondrial functions including electron transport chain, tricarboxylic acid cycle enzymes, and protein import, but, more importantly, novel proteins of unknown function were discovered amongst the subset of the most abundant proteins.

The concept of “global” profiling of the proteome is not achieved by 2D IEF/SDS-PAGE alone, because only the most abundant soluble proteins or peripheral membrane proteins are typically presented for counting and analysis. Lowly expressed proteins, hydrophobic proteins, and proteins with extreme pIs are poorly characterized using this classic approach. To overcome some of these limitations, a subsequent complementary study using direct analysis of whole mitochondrial tryptic lysates by LC–MS/MS was undertaken. The study confirmed 85% of proteins previously identified by gel-based approaches but also allowed identification of more than 200 new proteins [14]. Further studies have extended the known proteome by focusing on the import apparatus [15], the electron transport chain complexes [16, 17], metabolite carrier proteins [18], thioredoxin-regulated proteins [19], or ATP-binding proteins [20]. Currently over 550 proteins have been identified in the *Arabidopsis* mitochondrial proteome by direct identification through MS [21]. Smaller surveys of the mitochondrial proteome exist for rice [22], pea [23], and maize [24].

The Dynamic Mitochondrial Proteome: Tissue-Type Differences and Response to Development and Stress

Changes in mitochondrial function are underlined by changes in the proteome content in a qualitative or quantitative fashion. The differences in the appearance of the mitochondrial proteome on gels from different plant tissues have been known for over a decade [25, 26], but only recently were more systematically analyzed in pea to provide a list of tissue-specific mitochondrial proteins [23]. The development of the mitochondrial proteome during seed germination was highlighted in maize [27] and studied in detail in rice [28]. These studies showed the building of the tricarboxylic acid cycle and changes in the abundance of the import apparatus from immature promitochondrial structures to mature functional mitochondria.

Plants are highly susceptible to oxidative damage brought about by constant environmental stresses. In particular, the generation of ROS as a by-product of cellular respiration enhances the propensity of the oxidative injury to the mitochondrion.

Differential analysis of the mitochondrial proteome was applied to the study of *Arabidopsis* mitochondria subjected to different oxidative stresses, as well as pea plants subjected to drought, cold stress, and herbicide treatments to mimic environmental challenges [29, 30]. Changes in detoxification mechanisms (as determined by proteins with increased abundance), together with proteins prone to oxidative damage (as judged by decreased abundance or the presence of breakdown products), provided insights into the dynamics of plant mitochondria in response to oxidative stress. The targets of oxidative modification in plant mitochondria have also been studied by identification of proteins that have undergone carbonylation [31]. A recurring outcome of these studies indicates the presence of a subset of proteins with enhanced susceptibility to oxidative damage, exemplifying the potential of proteomics toward elucidating the metabolic consequences of stress in plant mitochondria.

25.3 METHODOLOGY AND STRATEGY

Isolation of *Arabidopsis* Mitochondria

Mitochondria can be isolated from many plant tissues, but *Arabidopsis* cell suspension cultures have become a tissue of choice due to the relatively homogeneous tissue, the production of large amounts of plant material in a short time period, and the resources for proteomic analysis in *Arabidopsis*. All steps provided in the methodology should be performed at 4°C to minimize protein degradation.

Briefly, contamination-free 7-day-old cell suspension cultures are harvested by filtration through one layer of muslin cloth. The cells are homogenized in batches of 50 g FW by 3 successive 15-s bursts in a pre-cooled Warning blender in 150 mL HB comprising of 45 mM mannitol, 50 mM tetra-sodium pyrophosphate, 0.5% (w/v) PVPP 40, 2 mM EGTA, 0.5% (w/v) BSA, 20 mM cysteine, pH 8.0. To prevent excessive heating and frothing, 30-s rest intervals are applied between blending. The homogenate is filtered through three layers of Miracloth and one layer of muslin to remove cell debris. The filtrate is then centrifuged at 1500g to pellet residual cell debris. The supernatant is retained and centrifuged at 24,000g for 15 min, forming a crude organelle pellet.

The crude pellet is resuspended in a small volume of wash buffer (comprising 0.3 M mannitol, 10 mM TES-NaOH, pH 7.5, and 0.1% (w/v) BSA). The solution is then layered over a step Percoll gradient. The gradient consists of 5 mL 40% Percoll at the bottom, overlaid with 20 mL 25% Percoll, which is then overlaid with 5 mL 18% Percoll, all prepared in wash medium. Following centrifugation at 40,000g for 30 min, a mitochondrial-enriched band is formed at the 25% and 40% Percoll interface. Enzymatic marker assays reveal the presence of plastidic and peroxisomal enzymes but not cytosolic contamination [12]. This band can be carefully aspirated, placed in new centrifuge tubes, and diluted in 5 times the volume with wash buffer for centrifugation at 24,000g for 15 min. The resulting pellet is overlaid onto a self-forming 35% Percoll gradient made in buffer containing 0.3 M sucrose, 10 mM TES-NaOH, pH 7.5, and 0.1% (w/v) BSA and centrifuged at 40,000g. The purified mitochondrial band is near the top of the gradient, while peroxisomal and plastidic contaminants are

located lower in the gradient. The mitochondrial band can be carefully aspirated and transferred to new tubes, where it is washed in 5 times the volume in wash buffer without BSA by centrifugation at 24,000g for 15 min. The supernatant is discarded and the wash process is repeated. Typically, a yield of 10 mg mitochondrial protein is obtained from 100-g fresh weight of starting material. This is sufficient for a significant number of analyses of mitochondrial function and protein composition.

Determination of Mitochondrial Purity

A pure mitochondrial preparation is biologically unattainable due to the dynamicity of interaction between intracellular compartments as well as the technical limitations of separating cell constituents by differences in size and density. Nevertheless, a necessity of proteomic analyses of subcellular organelles is samples of high purity to minimize identification of proteins that turn out to be the abundant protein profile of contaminating organelles. Marker enzyme assays can be used in assessing the contamination of isolated mitochondria. Typically, the abundance of peroxisomes can be identified by catalase, hydroxy-pyruvate reductase, or glyoxylate oxidase activities. Cytosolic contamination can be assayed by alcohol or lactate dehydrogenase (ADH/LDH) activities, while plastids can be identified by alkaline pyrophosphatase activity in dark-grown tissues, or by RuBisCO or chlorophyll content in photosynthetic tissues. On average, a 2 Percoll gradient method yields mitochondrial samples of 90–95% purity. The advent of free flow electrophoresis has increased the purity of the yeast mitochondrial preparations to 98% [32] and holds the promise of improving the depth of analysis of plant mitochondrial proteomes.

Subfractionation of Mitochondria

The subfractionation of mitochondria into the four basic compartments involves a combination of osmotic shock and differential centrifugation. Isolated mitochondria that have not previously been frozen are suspended in a low osmotic strength solution (0–50 mM sucrose, 2 mM EDTA, 10 mM MOPS, pH 7.2). The solution is gently stirred on ice for 15 min to osmotically shock mitochondria. The molarity of the solution then needs to be increased to 0.3 M sucrose by the slow addition with stirring of 2 M sucrose, 2 mM EDTA, 10 mM, pH 7.2 stock solutions. The solution is then centrifuged at 15,000g for 15 min at 4°C to pellet the mitoplasts comprising the inner membrane and matrix (IM and MA). The mitoplast may be stored frozen for fractionation at a later stage. The supernatant containing the outer membrane (OM) and the intermembrane space (IMS) is transferred to clean tubes and centrifuged at 200,000g for 90 min. The resulting supernatant contains the IMS and is removed and stored frozen. The crude OM pellet requires further purification to remove contaminating IM. To do this, the pellet is resuspended in a small volume of 100 mM Tris-HCl pH 7.4 and layered over a 4 mL 0.6 M/0.9 M step sucrose gradient. The gradient is then centrifuged in a swing-out rotor at 50,000g for 60 min. The OM band is collected from the 0.6 M/0.9 M interface and is only visible when performing fractionations using more than 100 mg of mitochondrial protein. Other visible bands are likely to be contaminants that should be discarded.

To separate the IM from the MA, the mitoplast pellet from the first centrifugation is diluted to 15 mg/mL low osmotic strength buffer and sonicated for 3×5 bursts with 20-s rest intervals on ice. The IM fraction is pelleted from the MA fraction by centrifugation at 80,000g for 60 min. Subcompartment sample purity can be assayed using cytochrome *c* oxidase for IM, fumarase for MA and antimycin A-insensitive NADH: cytochrome *c* oxidoreductase for OM. On average, enzymatic determination of cross-contamination from the compartments is less than 15%.

Preparation of Mitochondrial Membrane Protein Fractions

To minimize sample handling and maximize yield, a widely preferred method is the preparation of total mitochondrial membrane. Mitochondria are resuspended in MQ water at a final protein concentration of 2 mg/mL and subjected to repeated freeze–thaw cycles to lyse the mitochondria, releasing the soluble proteins. The lysate is centrifuged at 20,000g for 10 min, yielding a membrane pellet while soluble proteins remain in the supernatant. The freeze–thaw process is repeated to ensure thorough removal of soluble proteins. The membrane can be further fractionated to provide integral membrane proteins and peripheral membrane proteins. A combination of salt and alkaline treatment of membranes to electrostatically displace peripherally attached proteins is commonly performed using 0.1 M Na_2CO_3 , pH 11.5 [33]. Pelleted mitochondrial membrane fractions are resuspended in ice-cold 0.1 M Na_2CO_3 , pH 11.5 to a final protein concentration of 0.5 mg/mL. Gentle vortexing is applied for 30 s to resuspend the pellet and to facilitate peripheral protein detachment. The solution is centrifuged at 20,000g for 10 min. The resulting supernatant of peripheral proteins is then removed and treatment is repeated to ensure depletion of peripheral proteins. Other techniques for extraction of integral proteins from total membrane proteins using salt (0.5 M NaCl), alkaline (0.1 M NaOH), or C/M treatment have also been described [34].

Gel-Based Resolution of Integral Membrane Proteins

Tricine Double SDS-PAGE. The tricine dSDS-PAGE technique was developed to identify highly hydrophobic protein of the bovine heart mitochondria respirasome where hydrophobic proteins are predictably displaced above the diagonal [35]. Here we detail a protocol adapted from the original study and used to separate proteins from *Arabidopsis* mitochondrial integral membrane samples (Figure 25.1A). The first dimension separating gel consists of 1 M Tris-HCl, pH 8.45, 0.1% (w/v) SDS, 10% (w/v) acrylamide, 0.3% (w/v) methylenebisacrylamide and 6 M urea using AMPS and TEMED to initiate polymerization. The inclusion of 6 M urea results in greater dispersion of protein spots from the diagonal [35]. The stacking gel, made of 1 M Tris-HCl, pH 8.45, 0.1% (w/v) SDS, 5% (w/v) acrylamide, and 0.15% (w/v) methylenebisacrylamide, is overlaid onto the separating gel. Extracted integral membrane proteins are resolubilized in SDS-PAGE sample buffer before sample loading. For both dimensions, the electrophoretic buffers used are the cathode buffer (0.1 M Tris-base, pH 8.25 unadjusted, 0.1 M tricine, 0.1% (w/v) SDS) in the upper chamber and the anode buffer (0.2 M Tris-HCl, pH 8.9) in the lower chamber.

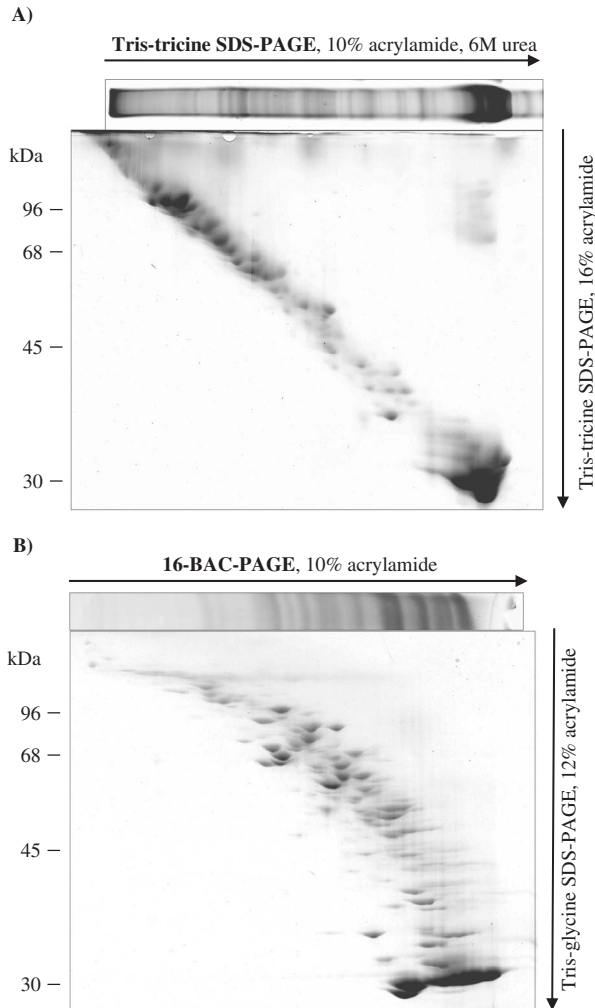


FIGURE 25.1. Comparison of tricine dSDS-PAGE and 2D 16-BAC PAGE/SDS-PAGE in the resolution of Na_2CO_3 -treated *Arabidopsis* mitochondrial membrane fractions. **(A)**, Tricine dSDS-PAGE analysis of 200 μg of proteins; **(B)**, 16-BAC PAGE/SDS-PAGE of 400 μg of proteins. Proteins were detected by colloidal CBB staining. In both cases, resolution of proteins above 30 kDa was targeted.

Following the first dimension electrophoresis, lanes are excised and acidified in 100 mM Tris-HCl, pH 2.2 for 30 min. For reasons yet to be clarified, acidification retards the movement of hydrophobic proteins in the second dimension causing displacement above the diagonal. The gel strips are treated in SDS-PAGE sample buffer for 5 min, and then they are rinsed briefly with water prior to application to the second dimension. The second dimension separating gel consists of 1 M Tris-HCl, pH 8.45,

0.1% (w/v) SDS, 16% (w/v) acrylamide, and 0.5% (w/v) methylenebisacrylamide. The stacking gel made of 1 M Tris-HCl, pH 8.45, 0.1% (w/v) SDS, 9.75% (w/v) acrylamide, and 0.3% (w/v) methylenebisacrylamide is overlaid onto the separating gel. Electrophoretic conditions are carried out as per the first dimension.

2D 16-BAC PAGE/SDS-PAGE. The acidic 16-BAC PAGE technique, using the cationic detergent benzyldimethyl-*n*-hexadecylammonium chloride (16-BAC), was developed for the high loading capacity and high resolution of complex protein mixtures, but was later adapted for its ability to separate integral membrane proteins [36, 37]. Described herein is a modification of the original protocol used to analyze *Arabidopsis* mitochondria integral membrane protein samples (Figure 25.1B). The first dimension separating gel consists of 75 mM KH₂PO₄, pH 2.1, 3 M urea, 10% (w/v) acrylamide, 0.26% (w/v) methylenebisacrylamide, 2.5 mM 16-BAC, 4 mM ascorbate, and 80 mM FeSO₄. Gel polymerization is initiated via the Fenton reaction following the addition of 0.025% H₂O₂. Overlaid on top is the stacking gel made of 125 mM KH₂PO₄, pH 4.1, 1.67 M urea, 4% (w/v) acrylamide, 0.34% (w/v) methylenebisacrylamide, 1.75 mM 16-BAC, 2 mM ascorbate, and 4.25 μ M FeSO₄.

Protein samples are resuspended in a twofold concentrated 16-BAC PAGE sample buffer containing of 7.5 M urea, 10% (w/v) 16-BAC, 10% (w/v) glycerol, 0.005% (w/v) pyronin Y, and freshly prepared 0.15 M DTT. The sample buffer should be heated to 60°C to prevent solidification. The electrode buffer consists of 2.5 mM 16-BAC, 150 mM glycine, and 500 mM H₃PO₄ with no pH adjustment. Electrophoresis is carried out toward the cathode (reverse polarity) until the tracking dye near the bottom of the gel. Following electrophoresis, the lanes of interest are excised and equilibrated in a buffer containing 100 mM Tris-HCl, pH 6.8 for 60 min. The lanes are then incubated with SDS-PAGE sample buffer for 5 min and applied to a standard SDS-PAGE for second dimension separation.

Non-Gel-Based LC-Based Resolution of Integral Membrane Proteins

RP-HPLC-MS/MS. Heazlewood et al. [14] performed complex analysis of *Arabidopsis* whole mitochondrial protein extracts. Batches of 500 μ g of mitochondrial protein were acetone precipitated and digested in 100 mM Tris-HCl, pH 8.5 overnight at 37°C with trypsin 1:10 (w/w). Soluble peptides from 15–20 μ g of this material were analyzed on a QStar Pulsar MS/MS system (Applied Biosystems, Foster City, CA) using an in-line Agilent 1100 capillary LC system incorporating a 0.150-mm Zorbax C18 reverse-phase column (Agilent, Palo Alto, CA). Peptides were analyzed by MS over a 10-h elution period with increasing acetonitrile concentrations from 2% to 80% (v/v) in water and 0.1% (v/v) formic acid at 4 μ L/min. Ions were selected for CID from an initial time-of-flight scan (1 s) if they occurred between 400 and 1500 atomic mass units (*m/z* ratio), had a charge series of MH₂, MH₃, or MH₄, and had an ion count > 10 cps. CID spectra were accumulated for 5 s. The resulting MS/MS-derived spectra were analyzed using Pro ID (Applied Biosystems) at error tolerances of +0.15 for MS and +0.05 for MS/MS, and only the top CID match for each spectrum was

accepted. The matching procedure initially used the calculated parent mass of an ion to screen the database for theoretical peptides that fell within the MS error tolerance range. Using this procedure, five independent mitochondrial isolations were digested and analyzed by LC–MS/MS. Proteins qualified for inclusion in the experimental set if the protein matched with a Pro ID confidence score of 98% or greater and if the protein was identified in two or more of the five experiments.

Brugiere et al. [34] performed complex analysis of *Arabidopsis* mitochondrial hydrophobic protein extracts, but these proteins samples were derived from running each sample on an SDS-PAGE in order to embed the proteins in a tight band of a polyacrylamide matrix to facilitate trypsination by a standard in-gel digestion protocol. Following digestion, peptides were extracted from gel pieces with 5% (v/v) formic acid solution and acetonitrile. After drying, tryptic peptides were resuspended in 0.5% aqueous TFA. The samples were injected into an LC-Packings (Dionex) nanoLC system eluted from a C18 column (75 $\mu\text{m} \times 150\text{ mm}$) using a gradient from solution A (5% acetonitrile: 95% water: 0.1% formic acid) to solution B (95% acetonitrile: 5% water: 0.1% formic acid) over 60 min at a flow rate of 200 nL/min. The LC system was directly coupled to a Q–TOF, and MS and MS/MS data were acquired and processed automatically. Samples from each treatment (C/M extraction, NaOH, and NaCl treatments) were individually analyzed by LC–MS/MS, while a separate sample retrieved from NaCl treatment was analyzed by a 2D chromatography approach using ion-exchange step elutions by 0, 100, 200, and 400 mM KCl upstream to reverse-phase C18 separation. Database searching was carried out using MASCOT. Proteins that were identified with at least two peptides showing a score higher than 40 were validated without any manual validation. For proteins identified by only one peptide having a score higher than 40, the peptide sequence was checked manually. Peptides with scores higher than 20 and lower than 40 were systematically checked and/or interpreted manually to confirm or cancel the MASCOT claim.

The main differences between these studies was the purification of a specific hydrophobic protein samples in the Brugiere et al. [34] study and the longer time used in analysis and the biological replication in the Heazlewood et al. [34] study (1 h versus 10 h and 1 replicate versus 5). These approaches identified 69 proteins in common, with 44 other proteins unique to the study by Brugiere et al. [34] and 347 unique to the study of Heazlewood et al. [14]. Many of the common set are *bonafide* membrane proteins from *Arabidopsis* mitochondria.

25.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Plant Mitochondria Hydrophobic Proteome: The Story So Far

The proteomics characterization of hydrophobic proteins has traditionally been a difficult task owing to problems with the low abundance and insolubility of this protein subset. As a consequence, the lack of attention with regard to the roles of organellar membrane proteins has limited progress toward understanding the intricate networks of interorganellar trafficking. To this end, the central role of the mitochondrion in bioenergetics makes it an excellent model organelle for such research.

The series of IEF/SDS-PAGE studies in *Arabidopsis* have identified 114 nonredundant proteins, but an analysis of the GRAVY score of these proteins reveals that only 5% of the total had a positive GRAVY score, and none were over the highly hydrophobic threshold of +0.3 (Table 25.1). This is an expected outcome because proteins with GRAVY scores greater than +0.3 are typically absent from IEF/SDS-PAGE profiles. In contrast, the larger set of 416 proteins from whole mitochondrial LC–MS/MS contained over 302 proteins not found by the IEF/SDS-PAGE sets, and 44 of these new proteins had positive GRAVY score, with 9 being over the +0.3 threshold (Table 25.1).

Millar and Heazlewood [18] presented a targeted study of the mitochondrial carrier family (MCF) which are putatively but not exclusively localised in the mitochondria. The MCF are a family of highly basic proteins made up of approximately 350 amino acids forming a conserved tripartite repeat structure with 6 putative TMDs and a mitochondrial energy transfer signal [18]. Based on these characteristics, genomic analysis predicted 45 putative members of the MCF in *Arabidopsis*, but this list was later revised to 58 putative members based on a widening of the criteria [18, 38]. The GRAVY scores of MCF proteins range from –0.23 to +0.34, but it can be skewed toward a negative value due to failure to take into account the removal of the amphiphilic mitochondrial targeting sequence to form the mature protein. In light of the *in silico* data, gel plugs were excised from the 28- to 38-kDa region from an SDS-PAGE for MS. Only two distinct adenine nucleotide transporters (At3g08580, At5g13490), a dicarboxylate/tricarboxylate carrier (At5g19760), an uncoupling protein (At3g54110), a phosphate carrier (At5g14040), and an unknown carrier protein (At4g01100) were identified. Not surprisingly, all but the unknown carrier have central roles in carrying out fundamental mitochondrial functions. The transport of metabolites, inorganic ions, and other small biomolecules continues to be defined in plant mitochondria, with recent discoveries of carriers for basic amino acids [39] and for AdoMet [40].

TABLE 25.1. Average Hydrophobicity of the Amino Acid Sequences of *Arabidopsis* Mitochondrial Proteins that Have Been identified by Different Methodologies^a

GRAVY	IEF/SDS-PAGE	New from Whole Mitochondria LC–MS/MS	New from Targeted Mitochondria LC–MS/MS
<–0.3	36	123	12
–0.3 to 0.0	74	135	17
0.0 to +0.3	8	35	9
>+0.3	0	9	6
Total	118	302	44

^aThe IEF/SDS-PAGE set of 114 is derived from Millar et al. [12], Krufft et al. [13], Sweetlove et al. [29] and Taylor et al. [47]; the new set from whole mt LC–MS/MS is the subset from Heazlewood et al. [14] that is not found in the IEF/SDS-PAGE set, and the new set from targeted LC–MS/MS is a subset of Brugiere et al. [34] that is not found in either of the other two sets. GRAVY (grand average of hydropathicity) values were calculated and recovered from the SUBA database (<http://www.suba.bcs.uwa.edu.au>), and are presented in ranges [21].

Brugiere et al. [34] sought to provide a more comprehensive survey of the *Arabidopsis* mitochondrial hydrophobic proteome using salt, alkali, and organic solvent extraction methods followed by 2D LC–MS/MS analysis. All but the unknown carrier protein from the targeted MCF study was identified. In addition, a lowly expressed carnitine/acylcarnitine carrier-like protein (At5g46800) was identified. A total of 114 proteins were identified, with 40% of these claimed to be previously undiscovered. Analysis of this set shows that 44 proteins of the Brugiere et al. [34] set were not present in the study of the whole mitochondria by similar non-gel methodologies, but that only 16 had positive GRAVY scores, and of these only 6 had GRAVY scores over +0.3 (Table 25.1).

Improving the Analysis of Plant Mitochondria Hydrophobic Proteome

Although studies have been scarce, invaluable information regarding the strengths and weaknesses of the research undertaken allows for the opportunity to develop well-informed approaches toward a more comprehensive survey of the plant mitochondrial hydrophobic proteome. Furthermore, a marrying of data-independent shotgun MS approaches with more targeted approaches using a variety of techniques will be required to search for proteins suspected to be part of the hydrophobic proteome. In Figure 25.2, a summary of the techniques currently used is provided. In addition, proposed strategies for improving the analysis of the plant mitochondrial hydrophobic proteome are illustrated. These proposed strategies are discussed below.

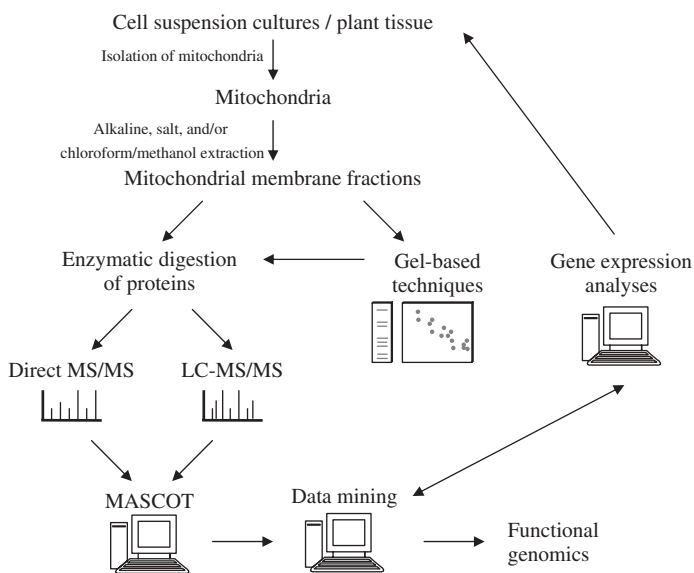


FIGURE 25.2. Flow diagram depicting current and proposed strategies to characterize the plant mitochondrial hydrophobic proteome.

Targeted Proteomics by Data Mining: Expression Analyses. The concentration of effort on *Arabidopsis* cell suspension cultures as a model for proteomics is justified by the relatively rapid generation of plant material, relative homogeneity, and need for a foundation of information from a single source. Nevertheless, there is also a need to address tissue-specific and stress-induced gene expression by means of more targeted approaches. For example, Genevestigator, a publicly available resource, can be interrogated to determine the spatiotemporal expression patterns of genes of interest [41]. Another useful resource is a series of databases called the AMPL (<http://www.cbs.umn.edu/arabidopsis/>) and the RMPL (<http://www.cbs.umn.edu/rice/>), which cluster membrane proteins into families [42]. Other major features of this database series are the HMMTOP topology predictions with putative TM peptides highlighted in the FASTA sequences. Unique to the AMPL is the compilation of NASCArray data providing developmental, tissue-specific, and abiotic stress-induced expression for these membrane protein-encoding genes. Moreover, transcript analyses provide insight into gene expression but also unveil clues to the functional significance of the protein of interest.

Proteins of interest can be analyzed using a multitude of online databases to determine putative topology, localization, and contamination. Commonly, the use of a single TM prediction or subcellular localization algorithm is liable to inaccurate data interpretations. For this reason, ARAMEMNON plant membrane protein database (<http://aramemnon.botanik.uni-koeln.de/>) was developed with a topology analysis section encompassing 17 TM prediction programs and 13 protein targeting programs to minimise bias [43]. Proteins that are putatively determined by topology prediction programs to not have TM regions may be artifacts of the extraction process. However, the possibility of the post-translational addition of membrane-anchoring hydrophobic moieties to the protein cannot be excluded. Protein contamination can be cross-checked against the *Arabidopsis* SUBA (<http://www.suba.bcs.uwa.edu.au>) [21].

Novel Techniques. The identification of low-abundant hydrophobic proteins is a heightened problem for mitochondria because the specialized electron transport and oxidative phosphorylation complexes represent a significant populace in the protein-enriched inner membrane. In addition, the 30-kDa carrier proteins are highly abundant such that SDS-PAGE fails to separate them from one another and potentially away from other lower-abundant proteins. To overcome this problem, the use of gel-based methods with high resolving capabilities or differential displacement is required. We have described two gel-based techniques that have been used to characterize other membrane systems because they had potential to provide high sample loading and resolving capacity. Figure 25.1 compares the resolving capabilities of the tricine dSDS-PAGE and 2D 16-BAC PAGE/SDS-PAGE in separating mitochondrial integral membrane proteins with emphasis on proteins with MWs greater than 30 kDa.

The primary advantage of using the tricine dSDS-PAGE for analyzing hydrophobic proteome is the specific differentiation of hydrophobic proteins above the diagonal. However, the resolving capability is unsuitable for highly complex samples such as the total mitochondrial membrane proteins. For tricine dSDS-PAGE, 29 spots can be visualized following staining with colloidal CBB, while greater than 90 spots

are reproducibly separated using 16-BAC PAGE/SDS-PAGE. The approximate three-fold increase in spot resolution makes 16-BAC PAGE/SDS-PAGE more effective for application to complex samples. More impressive is the resolution of specific highly abundant 30-kDa proteins from each other, particularly by 16-BAC PAGE alone. However, HMW proteins are poorly represented on the 2D 16-BAC PAGE/SDS-PAGE due to the formation of insoluble protein aggregates during sample suspension in 16-BAC PAGE sample buffer. Further investigations are warranted, but an analysis of the yeast mitochondrial hydrophobic proteome has shown the potential of this technique to resolve membrane proteins with high GRAVY scores and multiple TMDs [44].

Further development of non-gel LC–MS/MS analysis to focus on the hydrophobic proteome is required. This may include directed study of hydrophobic peptides by the capture and slow elution of peptides from different types of reverse-phase material. It will also require more systematic use of enzymes other than trypsin, because the Lys and Arg residues that dictate trypsin cleavage sites are rare in highly hydrophobic proteins and typically absent in membrane spanning regions. Minimal sample handling is also desired due to the intrinsic affinity of hydrophobic proteins to plastic surfaces. Furthermore, direct sample introduction, bypassing the need for LC-based peptide separations, may also decrease the loss of hydrophobic peptides that fail to elute from the hydrophobic stationary phases often used in this type of analysis.

25.5 CONCLUSIONS

The evolution of proteomics from the early maps of plant mitochondria to the systematic global studies and subsequent targeted proteomic studies of the *Arabidopsis* mitochondrial proteome has significantly enhanced our understanding of plant mitochondrial physiology. However, the application of hydrophobic proteomics toward elucidating the plant mitochondria integration into plant physiology is still in its infancy. Overcoming the challenges encountered in characterizing the dynamic plant mitochondrial hydrophobic proteome requires multidisciplinary but also targeted approaches. The ideas and techniques presented in this review are not exclusively intended for plant mitochondrial membranes, but are applicable to the study of other subcellular organelle membranes where the compartmentalization of plant cell metabolism by the membranes of organelles continues to be a subject of intrigue.

25.6 FIVE-YEAR VIEWPOINT

In the near future, the primary focus should be directed toward in-depth analysis of the mitochondrial proteome using an array of complementary techniques in both plant and cell culture models described in this chapter. The ongoing development of newer techniques and technologies may contribute toward a more in-depth analysis of the plant mitochondrial hydrophobic proteome, provided that there are no bottlenecks with access to the technology. For example, a method developed in yeast mitochondria known as “protein sequence tagging” has potential to improve the identification

of low-abundant hydrophobic proteins, but its limited mainstream availability is problematic for wide-scale implementation [45]. The importance of the need to convert concepts into reality also applies for the dissemination of data to the science community. PeptideAtlas, an online database that houses publicly available raw mass spectra data from a variety of organisms, enables researchers to mine data that may have been previously neglected [46]. However, the notion of publicly accessible raw mass spectra data is at present more conceptual than a reality for *Arabidopsis* research. Besides genomics and proteomics profiling, another key focus area is the initiation of functional genomics analysis of mitochondrial membrane-targeted gene products. The purpose of proteomics is not simply to generate exhaustive lists of proteins of gene products within an organelle but rather to provide clues about gene product localization and how this relates to organellar function and the impacts on plant cell metabolism. Hence, the priority would be investigating unique plant mitochondrial membrane proteins where putative functions cannot be extrapolated due to a lack of orthology to other eukaryotic proteins.

REFERENCES

1. Giege P, et al. 2003. *Plant Cell* **15**:2140–2151.
2. Bartoli CG, Pastori GM, and Foyer CH. 2000. *Plant Physiol* **123**:335–344.
3. Picciocchi A, Douce R, and Alban C. 2001. *Plant Physiol* **127**:1224–1233.
4. Rebeille F, et al. 1997. *EMBO J* **16**:947–957.
5. Raghavendra AS and Padmasree K. 2003. *Trends Plant Sci* **8**:546–553.
6. Moller IM. 2002. *Trends Plant Sci* **7**:235–237.
7. Vanlerberghe GC and McIntosh L. 1997. *Annu Rev Plant Physiol Plant Mol Biol* **48**:703–734.
8. Gray MW, Burger G, and Lang BF. 1999. *Science* **283**:1476–1481.
9. Rehling P, Pfanner N, and Meisinger C. 2003. *J Mol Biol* **326**:639–657.
10. von Heijne G. 1986. *EMBO J* **5**:1335–1342.
11. Emanuelsson O and von Heijne G. 2001. *Biochim Biophys Acta* **1541**:114–119.
12. Millar AH, et al. 2001. *Plant Physiol* **127**:1711–1727.
13. Krufft V, et al. 2001. *Plant Physiol* **127**:1694–1710.
14. Heazlewood JL, et al. 2004. *Plant Cell* **16**:241–256.
15. Lister R, et al. 2004. *Plant Physiol* **134**:777–789.
16. Eubel H, Jansch L, and Braun HP. 2003. *Plant Physiol* **133**:274–286.
17. Millar AH, et al. 2004. *Plant Mol Biol* **56**:77–90.
18. Millar AH and Heazlewood JL. 2003. *Plant Physiol* **131**:443–453.
19. Balmer Y, et al. 2004. *Proc Natl Acad Sci USA* **101**:2642–2647.
20. Ito J, Heazlewood JL, and Millar AH. 2006. *J Proteome Res* **5**:3459–3469.
21. Heazlewood JL, et al. 2005. *Plant Physiol* **139**:598–609.
22. Heazlewood JL, et al. 2003. *Plant Physiol* **132**:230–242.
23. Bardel J, et al. 2002. *Proteomics* **2**:880–898.
24. Hochholdinger F, Guo L, and Schnable PS. 2004. *Plant J* **37**:199–208.

25. Colas des Francs-Small C, et al. 1993. *Plant Physiol* **102**:1171–1177.
26. des Francs-Small CC, et al. 1992. *Plant Physiol* **98**:273–278.
27. Logan DC, et al. 2001. *Plant Physiol* **125**:662–672.
28. Howell KA, Millar AH, and Whelan J. 2006. *Plant Mol Biol* **60**:201–223.
29. Sweetlove LJ, et al. 2002. *Plant J* **32**:891–904.
30. Taylor NL, et al. 2005. *Mol Cell Proteomics* **4**:1122–1133.
31. Kristensen BK, et al. 2004. *Phytochemistry* **65**:1839–1851.
32. Zischka H, et al. 2003. *Proteomics* **3**:906–916.
33. Fujiki Y, et al. 1982. *J Cell Biol* **93**:103–110.
34. Brugiere S, et al. 2004. *Phytochemistry* **65**:1693–1707.
35. Rais I, Karas M, and Schagger H. 2004. *Proteomics* **4**:2567–2571.
36. Hartinger J, et al. 1996. *Anal Biochem* **240**:126–133.
37. Macfarlane DE. 1989. *Anal Biochem* **176**:457–463.
38. Picault N, et al. 2004. *Trends Plant Sci* **9**:138–146.
39. Hoyos ME, et al. 2003. *Plant J* **33**:1027–1035.
40. Palmieri L, et al. 2006. *Plant Physiol* **142**:855–865.
41. Zimmermann P, et al. 2004. *Plant Physiol* **136**:2621–2632.
42. Ward JM. 2001. *Bioinformatics* **17**:560–563.
43. Schwacke R, et al. 2003. *Plant Physiol* **131**:16–26.
44. Zahedi R, Meisinger C, and Sickmann A. 2005. *Proteomics* **5**:3581–3588.
45. Prinz T, et al. 2004. *J Proteome Res* **3**:1073–1081.
46. Desiere F, et al. 2005. *Genome Biol* **6**:R9.
47. Taylor NL, et al. 2004. *Plant Physiol* **134**:838–848.

PEROXISOME

Yuko Arai, Youichiro Fukao, Makoto Hayashi,
and Mikio Nishimura

26.1 INTRODUCTION

Peroxisomes are single membrane-bound organelles ubiquitously found in eukaryotic cells (Figure 26.1), in which a variety of oxidative metabolic reactions are compartmentalized. The functions differ with the species, developmental stage, and organ, even in the same organism. Scavenging of ROS, however, is a function common to mammals, yeasts, and plants [1–4]. In germinating seeds, fatty acid catabolism occurs by β -oxidation in the peroxisomes to acetyl-CoA and subsequent conversion of acetyl-CoA to succinate via the glyoxylate cycle [2]. These carbon skeletons provide energy before the seedlings develop the capacity to photosynthesize. During the post-germinative growth of seedlings, the peroxisomal function changes. In green plants, the peroxisomes contain enzymes of the glycolate pathway for the photorespiration process [5].

Peroxisomal enzymes are synthesized in the cytosol, and they function after their post-translational transport into peroxisomes. Most peroxisomal enzymes have been shown to contain one of two peroxisomal targeting signals (PTSs) within their amino

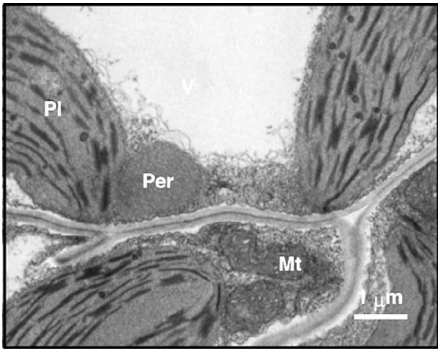


FIGURE 26.1. Electron microscopic observation of green cotyledon of *Arabidopsis*. Per, peroxisome; Mt, mitochondrion; Pl, plastid; V, vacuole.

acid sequences. The permissible combinations of tripeptide sequence for PTS1 are (C/A/S/P)-(K/R)-(I/L/M) at the carboxyl terminal [3]. Another type of targeting signal, PTS2, is a consensus sequence (A/L/Q/I)-X5-(H)-(L/I/F) (where X indicates any amino acid) at the amino terminal [3]. Using the PTS1 and the PTS2 sequences, we predicted that 286 peroxisomal genes are present in *A. thaliana* genome [4]. Transcriptomics analysis of all these genes revealed that plant peroxisomes are functionally differentiated into at least five classes, namely glyoxysomes, cotyledonary peroxisomes, leaf peroxisomes, root peroxisomes, and unspecialized peroxisomes, (Figure 26.2). Of these, functions of only 29 proteins have been demonstrated [4]. This result suggests that peroxisomes may have more diverse functions than those we know at present.

Recently, massive analyses have been done subsequent to the determination of the complete genome sequences of model organisms. The organelles are particular targets of these massive analyses, which include proteomics analyses [6]. Subcellular organelles represent attractive targets for proteomic analysis because they constitute

Organ	Seedling	Cotyledon Leaf		Root	Stem, Flower, Siliques
Classical nomenclature	Glyoxysome	Leaf peroxisome		Unspecialized peroxisome	
Novel nomenclature	Glyoxysome	Cotyledonary peroxisome	Leaf peroxisome	Root peroxisome	Unspecialized peroxisome
Specific genes	Glyoxylate cycle β-Oxidation	Glycolate cycle		Root specific	
		Cotyledon specific			
Constitutive genes	β-Oxidation Branched chain amino acids degradation			H ₂ O ₂ degradation Unknown	

FIGURE 26.2. Organ-specific differentiation of peroxisomes in higher plants.

distinct functional units, and the complexity of their protein composition is low relative to whole cells. Transcript and protein levels are not always well-correlated, and many organellar proteins function after transport into each organelle. Therefore, proteomic analysis of organelles should identify directly their functional proteins. In higher plants, proteomics analyses of peroxisomes, mitochondria, vacuoles, and chloroplasts have been reported following completion of the genome sequences of *Arabidopsis* [7–14].

To explore unidentified peroxisomal functions, we have been characterizing proteins in peroxisomes by proteomics analysis [7, 8]. Although we succeeded in characterizing novel peroxisomal proteins, we failed to identify some major peroxisomal proteins. This could be because the peroxisomal proteins detected by this method were limited due to the small quantity of sample. In this chapter, we show a method for proteomics analysis of *Arabidopsis* peroxisomes. Furthermore, to comprehensively analyze the peroxisomal proteins, we established an optimized method for isolating highly purified soybean peroxisomes from etiolated soybean cotyledons; these peroxisomal proteins were identified by proteomics analysis.

26.2 BRIEF BIBLIOGRAPHIC REVIEW

Peroxisomal Proteomics of Plants

We have been characterizing *Arabidopsis* peroxisomal proteins by PMF analysis coupled with 2-DGE and MALDI–TOF–MS [7, 8]. Recently, the peroxisomal localization of the smHSP, AtAc31.2, was further confirmed by isolation of leaf peroxisomes from *Arabidopsis* by two successive sucrose density gradients, protein separation by 1- and 2-DGE, and MS protein identification with TOF/TOF instrument [15].

Peroxisomal Proteomics of Yeast and Mammals

Rat liver peroxisomes isolated by cell fractionation were further purified by immun isolation using a specific antibody raised against a peroxisomal membrane protein, PMP70. The isolated peroxisomes were analyzed by 1D LC–MS [16]. PPIs in the rat liver peroxisomal membrane and matrix have been reported [17]. Male mice kidney peroxisomes were isolated and subjected to a combination of BN-PAGE and SDS-PAGE. The mouse kidney peroxisomal proteins were analyzed by MALDI–TOF–MS and Q–TOF–MS/MS analysis [18].

26.3 SPECIFIC METHODOLOGIES AND STRATEGIES

Selection of Plant Species for Peroxisomal Proteomics

Arabidopsis is a widely used model plant species whose genome is fully decoded and almost perfectly annotated (The *Arabidopsis* Genome Initiative, 2000), making

it very useful for proteomics analysis. To identify novel peroxisomal functions, we characterized peroxisomal proteins by 2-DGE with MALDI-TOF-MS [7, 8]. Five functionally identified proteins were detected from the peroxisomes of etiolated *Arabidopsis* cotyledons, but some major peroxisomal proteins other than APX, catalase, 3-ketoacyl-CoA thiolase (THI), isocitrate lyase (ICL), and MDH, were not identified [8]. This suggested limitations in the number of peroxisomal proteins that can be detected by this method.

The main limitation of proteomics analysis is the quantity of a sample. A large amount of protein obtained from highly purified sample is needed to identify minor proteins. However, it is very difficult to obtain a large amount of highly purified peroxisomes from *Arabidopsis* cotyledons. Although large cotyledons from plants such as cucumber, pumpkin, soybean, and sunflower may be suitable for isolation of peroxisomes, the complete sequence of the plant genome has not been determined in these species.

Arabidopsis sequences will not always provide positive hits in identification of proteins from other species. Recently, several investigators have suggested that EST databases from plants could be useful for identification of proteins [19]. The number of ESTs from soybean was larger than that from other plants with large cotyledons. There are 356,900 ESTs from soybean (*G. max*) in the NCBI EST database (May 19, 2006 update). Here, we report a method for isolating peroxisomes from *Arabidopsis* and soybean, using the relevant database to identify the proteins.

Isolation of Peroxisomes from *Arabidopsis* Cotyledons

Arabidopsis cotyledons (ecotype Columbia) were used for the isolation of peroxisomes. Figure 26.3 demonstrates the steps used to isolate peroxisomes by Percoll density gradient centrifugation from *Arabidopsis* cotyledons. All procedures were carried out at 4°C. Glyoxysomes were isolated from etiolated *Arabidopsis* cotyledons germinated at 22°C in the dark for 5 days. Leaf peroxisomes were isolated from *Arabidopsis* greening cotyledons germinated at 22°C in the dark for 5 days and were then transferred to light for 4 days. The *Arabidopsis* cotyledons (25 g fresh weight) were harvested from the seedlings and were homogenized with 100 mL of grinding buffer [20 mM pyrophosphate-HCl, pH 7.5, 1 mM EDTA, 0.3 M mannitol] in a chilled homogenizer (SMT COMPANY, Japan) for 10 s three times. The homogenate was squeezed through four layers of cheesecloth, and the residue was homogenized with another 50 mL of grinding buffer. The filtrates were combined and centrifuged at 1500g for 10 min to remove plastids and cell debris. After centrifugation at 10,000g for 20 min, the pellet was resuspended in 10 mL of grinding buffer and was centrifuged again. The resultant pellet was then resuspended in 4 mL of isolation buffer [10 mM HEPES-KOH, pH 7.2, 1 mM EDTM, 0.3 M mannitol] and was subjected to centrifugation in Percoll. The 4 mL suspended solution was layered directly on the top of 5 mL of 50% (v/v) and 30 mL of 28% (v/v) Percoll solutions [10 mM HEPES-KOH, pH 7.2, 1 mM EDTA, 0.3 M raffinose] and was centrifuged at 40000g for 40 min with slow acceleration and deceleration. After centrifugation, 1 mL fractions were fractionated (SJ-1211; ATTO, Japan).

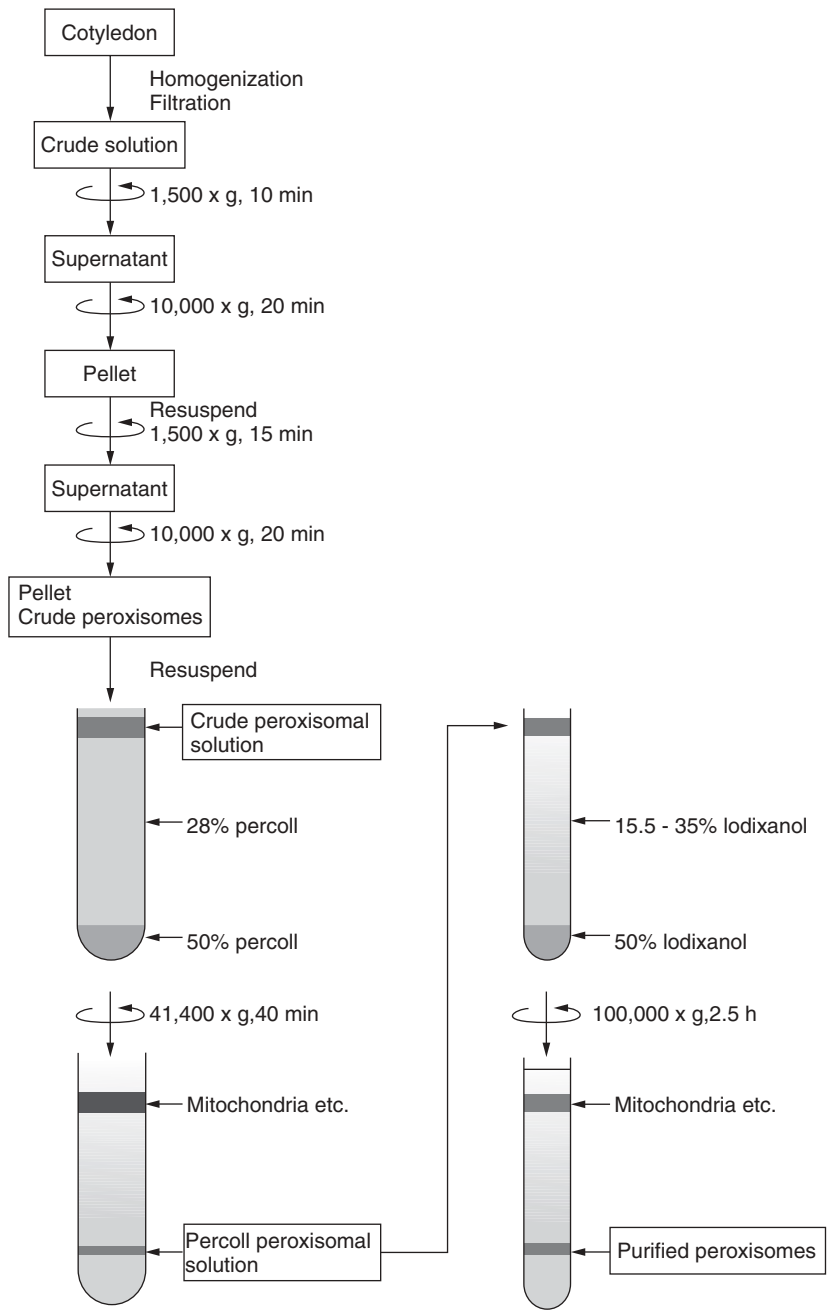


FIGURE 26.3. Flowchart for isolation of peroxisomes from *Arabidopsis* and soybean cotyledons. *Arabidopsis* peroxisomes were isolated by Percoll density gradient centrifugation. Soybean peroxisomes were isolated by Percoll and iodixanol density gradient centrifugation.

To confirm that the peroxisomal fraction was not contaminated with the proteins of other organelles, we measured enzyme activities and conducted immunoblot analyses. Activities of catalase [20] and cytochrome *c* oxidase (COX) [21] were measured as markers of peroxisomes and mitochondria, respectively. Immunoblot analysis was performed by the method of Towbin et al. [22]. Immunologic reactions were detected by monitoring the activity of HRP (ECL system; GE Healthcare BioSciences, Little Chalfont, UK). Anti-catalase [23, 24], anti-Cpn10 [25], anti-Cpn20 [26], anti-BiP [27], anti-RD21A [28], and anti-aconitase antibodies [29] were used to define the separated organelles in the gradient as markers of peroxisomes, mitochondria, plastids, ER, vacuoles, and cytosol, respectively. The peroxisomal fractions were centrifuged at 100,000g for 90 min. The concentrated peroxisomal fraction sediment near the bottom was collected. The isolated peroxisomal proteins were separated by 2-DGE using IPG strips of pH 3–10 or pH 6–11 and SDS-PAGE.

Isolation of Peroxisomes from Soybean Cotyledons

Soybean seeds (*G. max* cv. Bansei Shirodaizu) were used for isolation of peroxisomes. Figur 26.3 demonstrates the steps to isolate peroxisomes by Percoll and iodixanol density gradients centrifugation from soybean cotyledons. All procedures were carried out at 4°C. Glyoxysomes were isolated from etiolated soybean cotyledons germinated at 22°C in the dark for 7 days.

The soybean cotyledons (100 g fresh weight) were harvested, and then they were homogenized with 200 mL of grinding buffer in a chilled homogenizer twice for 3 s. The homogenate was squeezed through four layers of cheesecloth and the residue was homogenized with another 200 mL of grinding buffer. The filtrates were combined and centrifuged at 1500g for 10 min to remove plastids and cell debris. After centrifugation at 10,000g for 20 min, the pellet was resuspended in 150 mL of grinding buffer and was then centrifuged. The resultant pellet was then resuspended in 4 mL of isolation buffer and centrifuged in Percoll. Soybean peroxisomes were separated by Percoll density gradient centrifugation by the same method as that used for *Arabidopsis*. Peroxisomal fractions obtained from Percoll gradients were combined and diluted fivefold with isolation buffer, followed by centrifugation at 4800g for 10 min. The pellet obtained was resuspended in 2 mL of isolation buffer. This solution was carefully layered onto an iodixanol density gradient that consisted of 2 mL of 50% (w/v) iodixanol solution [25 mM sucrose, 2.5 mM MOPS-HCl pH 7.2, 0.5 mM EDTA, 0.05% ethanol] and 13.6 mL of a linear iodixanol density gradient from 36% (w/v) to 15.5% solution [0.3 M sucrose, 5 mM MOPS-HCl pH 7.2, 1 mM EDTA, 0.1% ethanol]. The gradient was then centrifuged at 100,000g for 2.5 h in a SW28.1 swinging rotor (Beckman Coulter, San Diego, CA). Peroxisomal fractions obtained from the iodixanol density gradient were collected. The fractions were analyzed by immunoblotting to confirm that the peroxisomal fraction was not contaminated with proteins of other organelles. Proteins of the peroxisomal fraction were collected by methanol/chloroform/water precipitation [30]. The isolated peroxisomal proteins were separated by 2-DGE using IPG strips of pH 3–10 or pH 6–11 and SDS-PAGE.

MALDI-TOF-MS and PMF Analyses

The spots were excised from the silver-stained gel using the method of Gharahdaghi et al. [31], and the proteins were digested with trypsin (Promega, Tokyo, Japan) and lysyl endopeptidase (Wako Pure Chemical, Osaka, Japan), respectively. The peptide mixtures were concentrated using an evaporator and purified by C18 Zip-Tip (Millipore, Billerica, MA). Each peptide solution was mixed with a solution of 2:1 0.01% TFA acetonitrile saturated with α -cyano-4-hydroxycinnamic acid and was then dried on the surface of a standard stainless steel target (Bruker Daltonik GmbH, Leipzig, Germany). MS analysis was conducted by MALDI-TOF-MS (Reflex III, Bruker Daltonik GmbH) in positive ion reflector mode. Calibration was carried out using a standard peptide mixture (Sigma-Aldrich, St. Louis, MO). The mass spectra data were collected from monoisotopic peaks falling in the m/z range of 500–4000 Da.

Arabidopsis proteins were identified by searching a protein sequence database (ProFound™ database at Rockefeller University). Soybean proteins were identified with the TIGR Soybean Gene Index database, using MASCOT software in the PMF search mode. The TIGR Soybean Gene Index database is a nonredundant EST database. The database listed soybean Tentative Consensus (TC) sequences were created by assembling ESTs into virtual transcripts. In some cases, the TCs contain full or partial cDNA sequences obtained by classical methods. The TCs contain information on the source library and abundance of ESTs and in many cases represent full-length transcripts. Alternative splice forms are built into separate TCs. The database was downloaded from “The TIGR Soybean Gene Index (September 20, 2004 released version 12.0)” in <http://www.tigr.org/tdb/tgi/> [32–35].

26.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Characterization of the Glyoxysomal Proteome of *Arabidopsis*

Glyoxysomes were isolated from etiolated *Arabidopsis* cotyledons by Percoll density gradient centrifugation. Fractions isolated on the Percoll density gradient were examined for catalase and COX activities, which are markers for glyoxysomes and mitochondria, respectively. The highest level of catalase activity was localized to fraction 29, whereas COX activity was barely detected (Figure 26.4A). Immunoblot analysis with anti-catalase antibody revealed that glyoxysomal proteins occur mainly in fraction 29, together with catalase activity (Figure 26.4B). No marker proteins of other organelles were found in fraction 29 (Figure 26.4B). We concluded, therefore, that the glyoxysomes were prepared at high purity and in high concentrations.

The glyoxysomal proteins separated by 2-DGE were visualized by silver staining. A representative 2D gel protein pattern of *Arabidopsis* glyoxysome is presented in Figure 26.5. For qualitative analysis, the proteins were separated by 2-DGE. A total of 49 protein spots were excised from 2D gels and digested with the trypsin. The peptide fragments were extracted and analyzed by MALDI-TOF-MS. We could identify five glyoxysomal proteins on the 2D gels (Table 26.1). Among them, three proteins, THI, ICL, and MDH, were enzymes of the fatty-acid β -oxidation cycle and glyoxylate cycle, and two proteins, catalase and APX, were involved in scavenging H_2O_2 (Table 26.1).

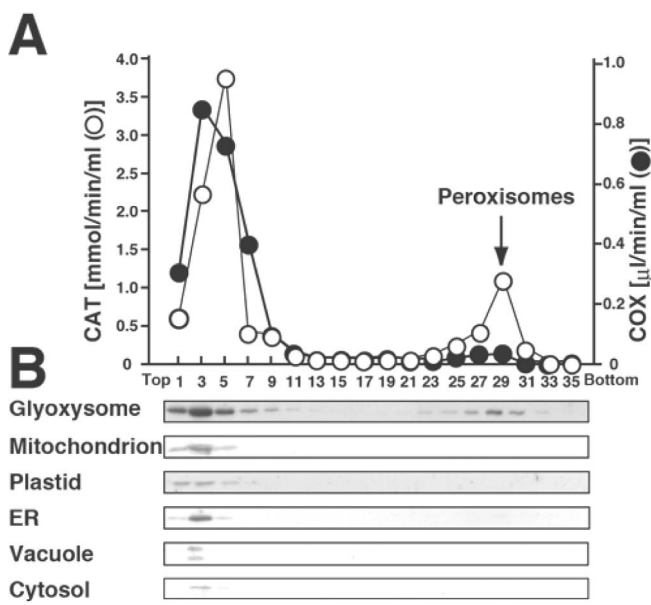


FIGURE 26.4. Isolation of glyoxysomes from etiolated *Arabidopsis* cotyledons by Percoll density gradient centrifugation. **(A)**, The distribution of marker enzyme activities on the Percoll density gradient. CAT and COX are markers of glyoxysomes and mitochondria, respectively. Fraction 1 represents the top of the gradient. **(B)**, Immunoblot of SDS-PAGE gel showing the level of catalase, Cpn10, Cpn20, BiP, RD21A, and aconitase, which are markers for glyoxysomes, mitochondria, plastid, ER, vacuoles, and cytosol, respectively.

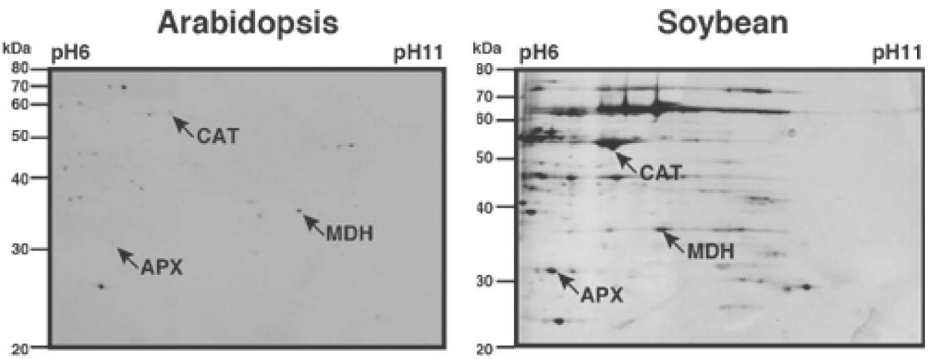


FIGURE 26.5. Two-dimensional map of peroxisomal proteins from etiolated cotyledons of *Arabidopsis* and soybean. Peroxisomal proteins were separated by 2-DGE with IEF in the first dimension and SDS-PAGE in the second dimension. Proteins were detected by silver nitrate staining. The numbers above the gel indicate pI values, and the numbers on the left indicate molecular masses of standard proteins. The arrows mark spots corresponding with some identified proteins. CAT, catalase.

TABLE 26.1. Identified Peroxisomal Proteins

Protein Name	Accession Number		
	Arabidopsis		Soybean
	Etiolated	Greening	Etiolated
β-Oxidation			
Acyl-CoA oxidase			TC216734
Multifunctional protein			TC216121
3-Ketoacyl-CoA thiolase	At2g33150		TC225397
Glyoxylate Cycle			
Malate synthase			TC225586
Isocitrate lyase	At3g21720		TC204772
Malate dehydrogenase 1	At2g22780		TC227709
Malate dehydrogenase 2			TC224978
Citrate synthase			TC227930
Glycolate pathway			
Hydroxypyruvate reductase		At1g68010	
Glutamate:glyoxylate aminotransferase 1		At1g23310	
Glutamate:glyoxylate aminotransferase 2		At1g70580	
H₂O₂ Degradation			
Catalase	At1g20620	At1g20620	TC204238
Ascorbate peroxidase	At4g35000	At4g35000	TC205158
Other			
Aspartate aminotransferase		At5g11520	TC204869

It has already been indicated that glyoxysomes contain enzymes of fatty-acid β-oxidation cycle and glyoxylate cycle [3]. Acyl-CoA oxidase (ACX), THI, and MFP, and which exhibits hydratase, dehydrogenase, epimerase, and isomerase activities are involved in the core pathway of the fatty-acid β-oxidation cycle. Malate synthase, ICL, MDH, and citrate synthase (CS) are involved in the glyoxylate cycle. However, we failed to identify major peroxisomal proteins, except for THI, ICL, and MDH.

Characterization of the Leaf Peroxisomal Proteome of *Arabidopsis*

Leaf peroxisomes were isolated from green *Arabidopsis* cotyledons by Percoll density gradient centrifugation. The leaf peroxisomal fraction of the Percoll density gradient was free of contamination from other organelles because no mitochondrial, plastidic, or ER proteins were detected in the analysis of the enzyme activities or immunoblotting.

The leaf peroxisomes proteins separated by 2-DGE were visualized by silver staining. For a qualitative survey, the proteins were separated by 2-DGE, and we excised a total of 53 protein spots from 2D gels and digested them with trypsin. The peptide fragments were extracted and analyzed by MALDI–TOF–MS. We could identify six leaf peroxisomal proteins on the 2D gels (Table 26.1). Among them, two proteins, hydroxypyruvate reductase (HPR), glutamate:glyoxylate aminotransferase (GGT) 1,

and GGT2, were related to the glycolate pathway, two proteins, catalase and APX, functioned to scavenge H_2O_2 (Table 26.1). HPR, glycolate oxidase, MDH, and some aminotransferases are involved in the glycolate pathway. However, we could not identify glycolate oxidase by this method.

Protein Identification Using Soybean EST Database

Soybean glyoxysomes were isolated from etiolated soybean cotyledons by Percoll and iodixanol density gradient centrifugation. The glyoxysomal fraction of the iodixanol density gradient was free of contamination from other organelles, since we could not detect mitochondrial, plastidic, and ER proteins after the immunoblot analyses (data not shown).

Glyoxysomal proteins separated by 2-DGE were visualized by silver staining. Figure 26.5 shows a representative 2D gel protein pattern of soybean glyoxysome. There were more protein spots than in the 2D gel of *Arabidopsis* glyoxysomal proteins (Figure 26.5). After separation by 2-DGE, we excised a total of 269 protein spots from the 2D gels and digested them with the lysyl endopeptidase. The peptide fragments were extracted and analyzed by MALDI-TOF-MS. We detected the spectra of 200 of these protein spots and assigned 71 TCs. We could identify most of the major glyoxysomal proteins from the 71 TCs (Table 26.1). Among them, three proteins, ACX, MFP and THI, were related to the fatty acid β -oxidation cycle, five proteins, MS, ICL, MDH1, MDH2, and CS, functioned in the glyoxylate cycle, and two proteins, catalase and APX, functioned to scavenge H_2O_2 (Table 26.1). The proteins identified from etiolated soybean cotyledons included most of the major enzymes for known glyoxysomal functions [41].

26.5 CONCLUSIONS

The requirements for proteomic analysis are a high-quality database and an adequate amount of sample. The quality of the proteomic analysis declines severely by the lack of either of them. The *Arabidopsis* database is very useful for proteomics analysis, because the genome is fully decoded and almost perfectly annotated. However, it is very difficult to obtain a large amount of highly purified peroxisomes from *Arabidopsis*. Therefore, few proteins related to peroxisomal functions have been identified by proteomics analysis in *Arabidopsis*. To enable a more comprehensive analysis of peroxisomal proteins, we performed proteomics analysis using soybean. We established a method for isolating highly purified peroxisomes from the cotyledons. The isolated peroxisomes contained enough protein for analysis by PMF coupled with 2-DGE and MALDI-TOF-MS. We identified more peroxisomal proteins by this method than by peroxisomal proteomics analysis in *Arabidopsis*. The identified proteins contained most of the major enzymes for glyoxysomal functions. Further analysis of the peroxisomal soybean proteins using proteomics analysis will increase our knowledge of the peroxisomal functions.

26.6 FIVE-YEAR VIEWPOINT

This improved success rate in performing PMF in this study was obtained by the use of soybean peroxisomes and a nonredundant EST database. However, the soybean EST database does not contain all soybean full-length ESTs. In the future, it will be most important to use databases with the full-length genomic sequences of plants from which peroxisomes can be isolated.

Identification of Novel Metabolic Reactions. Recent progress with mutant characterization and metabolite profiling has led to the identification of new peroxisomal genes and enzymes, elucidation of substrate specificities, definition of the pathway of carbon metabolism, degradation of branched amino acids, and the biosynthesis of plant hormones [36]. On the other hand, many putative enzymes were identified from *Arabidopsis* peroxisomes by bioinformatics-based strategy [4, 37]. Some peroxisomal proteins, homologous to prokaryotic enzymes of a degradation pathway for aromatic compounds, have been proposed to be involved in peroxisomal biosynthesis of aromatic and cyclic plant hormones such as JA, auxin, as well as SA [37]. However, physiological peroxisomal functions depend on the environmental conditions. To date, proteomics analyses have been reported only for etiolated and green cotyledons, as well as green leaves [7, 8, 15]. Protein profiles need to be obtained using peroxisomes isolated from various organs and various developmental stages in the future to identify metabolic reactions for novel functions in peroxisomes.

Analysis of Regulation Mechanisms of Peroxisomal Functions. Glyoxysomes in etiolated tissues are transformed directly into leaf peroxisomes during the greening process [38, 39]. The reverse transition of leaf peroxisomes to glyoxysomes is observed in senescencing tissues [40]. During these transitions, new peroxisomes do not appear and the initial enzyme sets are replaced by new ones [38, 39]. The search for proteases involved in the specific degradation of glyoxysomal or leaf peroxisomal proteins has not yet revealed unique proteases. Proteome analysis has been used to study changes in protein expression, modification, and degradation. Further analysis of peroxisomal proteins in various tissues using comparative proteomics analysis will provide us with molecular understanding of the mechanisms involved in transformation of peroxisomal function.

Combination of Proteomics with Transcriptomics Profiling. Many putative proteins of peroxisomes were identified from *Arabidopsis* using a bioinformatics-based strategy [4, 37]. Analyses of the peroxisomal gene expression profiles revealed that peroxisomal differentiation is caused by the expression of specific genes that are induced in specific organs in addition to constitutively expressed genes [4]. These data suggested that the functional state of plant peroxisomes is determined mainly at the transcription level. The profiles of proteomics data did not agree completely with those of transcriptomics data. Further analysis by transcriptomics and proteomics profiling will increase our knowledge of the regulation mechanism of peroxisomal function.

ACKNOWLEDGMENTS

We thank Ms. Maki Kondo (National Institute for Basic Biology) for the electron micrograph, Ms. Yumiko Makino and Ms. Tomoko Mori (National Institute for Basic Biology) for their technical support in the analysis of MALDI–TOF–MS, Ms. Hiroyo Nishide (National Institute for Basic Biology) for her technical support in the database search, and Dr. John T. Christeller (Horticulture and Food Research Institute of New Zealand) for discussions.

REFERENCES

1. Wanders RJ. 2004. *Am J Med Genet* **126A**:355–375.
2. Baker A, et al. 2006. *Trends Plant Sci* **11**:124–132.
3. Mano S and Nishimura M. 2005. *Vitam Horm* **72**:111–154.
4. Kamada T, et al. 2003. *Plant Cell Physiol* **44**:1275–1289.
5. Tolbert NE, et al. 1968. *J Biol Chem* **243**:5179–5184.
6. Taylor SW, Fahy E, and Ghosh SS. 2003. *Trends Biotechnol* **21**:82–88.
7. Fukao Y, Hayashi M, and Nishimura M. 2002. *Plant Cell Physiol* **43**:689–696.
8. Fukao Y, et al. 2003. *Plant Cell Physiol* **44**:1002–1012.
9. Kruft V, et al. 2001. *Plant Physiol* **127**:1711–1727.
10. Millar AH, et al. 2004. *Plant Mol Biol* **56**:77–90.
11. Schubert M, et al. 2002. *J Biol Chem* **277**:8354–8365.
12. Peltier JB, et al. 2000. *Plant Cell* **12**:319–342.
13. Peltier JB, et al. 2002. *Plant Cell* **14**:211–236.
14. Shimaoka T, et al. 2004. *Plant Cell Physiol* **45**:672–683.
15. Ma CL, et al. 2006. *Plant Physiol* **141**:47–60.
16. Kikuchi M, et al. 2004. *J Biol Chem* **279**:421–428.
17. Makkar RS, et al. 2006. *Arch Biochem Biophys* **451**:128–140.
18. Ofman R, et al. 2006. *Biochem J* **93**:537–543.
19. Lippert D, et al. 2005. *Proteomics* **5**:461–473.
20. Lück H. 1965. In: *Methods of Enzymatic Analysis* (Bergmeyer HU, ed.) Academic Press, New York, pp. 885–894.
21. Hodges TK and Leonard TR. 1974. *Methods Enzymol* **32**:392–406.
22. Towbin H, Staehelin T, and Gordon J. 1979. *Proc Natt Acad Sci USA* **76**:4350–4354.
23. Yamaguchi J and Nishimura M. 1984. *Plant Physiol* **74**:261–267.
24. Yamaguchi J, Nishimura M, and Akazawa T. 1984. *Proc Natt Acad Sci USA* **81**:4809–4813.
25. Koumoto Y, et al. 1996. *Plant J* **10**:1119–1125.
26. Koumoto Y, Shimada T, and Kondo M. 1999. *Plant J* **17**:467–477.
27. Hatano K, et al. 1997. *Plant Cell Physiol* **38**:344–351.
28. Yamada K, et al. 2001. *Plant Physiol* **127**:1626–1634.
29. Hayashi M, et al. 1995. *Plant Cell Physiol* **36**:669–680.
30. Wessel D and Flugge UI. 1984. *Anal Biochem* **138**:141–143.
31. Gharahdaghi F, et al. 1999. *Electrophoresis* **20**:601–605.

32. Quackenbush J, et al. 2001. *Nucleic Acids Res* **29**:159–164.
33. Lee Y, et al. 2002. *Genome Res* **12**:493–502.
34. Tsai J, et al. 2001. *Genome Biology* 2 Software 0002.1–0002.4.
35. Pertea G, et al. 2003. *Bioinformatics* **19**:651–652.
36. Hayashi M and Nishimura M. 2006. *Biochim Biophys Acta* **1763**:1382–1391.
37. Reumann S, et al. 2004. *Plant Physiol* **136**:2587–2608.
38. Nishimura M, et al. 1986. *Plant Physiol* **81**:313–316.
39. Titus DE and Becker WM. 1985. *J Cell Biol* **101**:1288–1299.
40. Nishimura M, et al. 1993. *Protoplasma* **175**:131–137.
41. Arai Y, et al. 2008. *Plant Cell Physiol*, in press.

UNRAVELING PLANT VACUOLES BY PROTEOMICS

Songqin Pan and Natasha Raikhel

27.1 INTRODUCTION

The vacuole of plant cells is essential to plant life. It plays many biological roles during plant development and responses to environmental changes. In mature plant vegetative tissues, such as the leaf, the central vacuole represents the largest cellular organelle and occupies more than 80% of the total cell volume. The reasons for the need of such a large vacuole by plant cells are still not entirely clear, although there is a consensus view that it is required to maintain the cellular turgor pressure, which serves as a “bone” to provide physical support for the proper cell shape. The importance of vacuoles can be also explained by the fact that it still stands as the giant in the “center” of a cell even after millions of years of evolution. Therefore, it is not surprising that perturbation in vacuole biogenesis is lethal to plants [1].

In contrast to yeast and animal cells that have only a single type of vacuole or lysosome, plants generally contain multiple types of vacuoles in a single cell. The lytic vacuole of plant cells is an acidic organelle and functionally equivalent to the yeast vacuole and animal lysosome [2]. The protein storage vacuole is of neutral pH and is thought to be plant-specific, where proteins such as lectins, chitinases, legumins, and

other metabolites are accumulated as the source of nutrients for plant development [3]. Plant vacuoles originate as the prevacuolar vesicles and expand in size through a series of events in enlargement and fusion [2]. The number, shape, and size of these prevacuoles may vary from cell to cell. In many different types of plant cells, the lytic vacuole and the protein storage vacuole may fuse to form a large central vacuole that retains functional characteristics of both parent compartments.

The plant vacuole is a multifunctional organelle. The major functions of the vacuole include maintenance of cellular homeostasis, storage, and digestion of proteins and metabolites, signaling in stress and defense responses, and sequestration of compounds that are toxic to plant cells [4]. Recent findings even suggest that plant cell vacuoles may be involved in signaling in tropic responses such as the gravitropism, which is a plant-specific function that does not appear to exist in yeast and animals [2]. To support these functions, vacuoles contain a large number of hydrolytic and biosynthetic enzymes, ion and metabolite transporters, among many others [5]. The molecular basis for these proteins being targeted and localized to plant vacuoles is an area of extensive research in vacuole biology.

In view of integrated cellular systems, the plant vacuole is part of the endomembrane system, which includes several other compartments such as the ER, Golgi apparatus, *trans*-Golgi network, and endosomes [2]. Transport between these cellular compartments is accomplished by vesicles. Although not completely understood, vesicle delivery and fusion from the *trans*-Golgi network represents the primary mechanism for the delivery of soluble and tonoplast proteins to the vacuole [2]. In general, soluble proteins targeted to vacuoles contain a specific sequence recognition motif at either the N- or C-terminus, termed amino-terminal propeptide (NTPP) and carboxy-terminal propeptide (CTPP), respectively. These sorting signals are recognized by cargo receptors such as the vacuolar sorting receptors (VSRs), and the proteins carrying the NTPP motif are directed to the clathrin-coated vesicles [2]. Fusion of these vesicles to the tonoplast involves the SNARE proteins. Little is known about mechanisms of delivery of proteins carrying the CTPP motif. In addition, it has become clear that vacuolar proteins can also be transported via other unknown mechanisms [6]. For a recent comprehensive review of the transport mechanisms to plant vacuoles, see an article by Surpin and Raikhel [2].

Our understanding of biological functions of plant vacuoles has greatly benefited from genetics studies in recent years. Phylogenetic analysis, genetic screening, and yeast complementation tests have identified a number of genes that are involved in vesicle trafficking and vacuole function in plants. For example, these approaches have been effective in the investigation of plant SNARE proteins that link vacuoles to several important cellular functions including gravitropism, autophagy, cytokinesis, and ABA-mediated signaling [2]. However, these genetic studies may be limited by the functional redundancy of gene families or by being lethal due to mutagenesis of a single essential gene. Proteomics approaches are able to avoid the cellular and genetic perturbations that are often associated with the genetics approaches. Vacuolar proteins can be directly analyzed qualitatively and quantitatively by proteomics methods with plant tissues or cells under physiological conditions. Determining the protein

constituents of plant vacuoles as well as their dynamic changes in abundance during development and responses to stress may provide a global view of this organelle regarding its biogenesis and function.

27.2 BIBLIOGRAPHIC REVIEW: ROLE OF PROTEOMICS IN UNDERSTANDING BIOLOGY OF PLANT VACUOLES

Proteomics studies of the plant vacuole are just emerging. There are only six research reports that are related to vacuoles of *Arabidopsis*, barley, and rice [6–11]. In these publications, the investigators tried to catalog all vacuolar proteins from either tonoplast or intact vegetative vacuoles. Although these studies represent the early stages of research in vacuolar proteomics, from the findings reported we have observed several interesting novel insights into vacuole biology.

Arabidopsis

The model plant *A. thaliana* has been the primary focus of proteomics studies of vacuoles [6, 8–10]. Along with others, we have attempted to characterize *Arabidopsis* vacuole proteins using MS-based proteomics approaches. In our study, vegetative vacuoles are isolated from mature leaves. Total vacuoles or fractioned tonoplasts are subjected to 2D LC/MS/MS or 1D SDS-PAGE coupled with LC–MS/MS analyses. With the multiple methods used, we are able to identify 402 proteins in total, which represents the most comprehensive set of the vacuolar proteins thus far [6]. Other groups have also investigated tonoplast proteins, and the number of identified proteins varies from 43 to 163 [8–10]. Bioinformatics analyses of all these proteins are able to assign the vacuolar proteins into six functional groups, including transporters, glycosidases, proteases, and proteins that are related to stress responses, membrane remodeling, and cytoskeletal elements [6]. The diverse functions of the vacuole suggest that a large array of proteins is required to conduct all of these processes. The discovery of many types of proteins in vacuoles is consistent with this view.

A surprising finding from our study is that the 42-kDa myrosinase-associated protein (MAP; At3g14210) is the most abundant protein of the leaf vacuole proteome [6]. When compared to other vacuole samples such as those from suspension-cultured cells, the strong staining of this protein on 1D SDS-PAGE appears leaf-specific, suggesting that leaf tissues require a higher amount of MAP in vacuoles than do other nondifferentiated cells. The functional role of MAP in vacuoles is unknown, although one possibility is that it may be important for maintaining the osmotic pressure.

Before these proteomics studies, other evidences suggested that CTPP and NTPP pathways in trafficking were functional in *Arabidopsis*. Translational fusions of the CTPPs from barley lectin or tobacco chitinase to other *Arabidopsis* secreted proteins were sufficient to redirect them from secretion to the vacuole location in *Arabidopsis* plants [12]. Then, however, no *Arabidopsis* endogenous proteins were known to contain a functional CTPP. Furthermore, only a single *Arabidopsis* protein (aleurain, At5g60360) was known to be transported *via* the NTPP pathway [13]. Understanding

molecular mechanisms of the CTPP and NTPP pathways in *Arabidopsis* awaited the discovery of those endogenous proteins that contain the CTPP or NTPP sorting signals. The global surveys of the vacuole proteome have thus played an important role in this regard. Our and other studies have revealed several CTPP-containing proteins including six POX and three jacalin-like lectins. In addition, 42 proteins that contain a putative NTPP signal are observed [6]. Nine proteins that contain the potential ER retention signals are also identified, which is consistent with the fact that ER bodies are involved in vesicle trafficking to the vacuole. These findings have clearly demonstrated the advantage of proteomics approaches over other conventional methods, with which only a single gene or protein can be characterized at a time.

The finding of a specific SNARE-pin complex at the tonoplast by proteomics is another important observation. Previous studies on the SNARE-pin composition at the tonoplast and the prevacuolar compartment are somewhat controversial. Some suggested that it might be localized to the tonoplast [14], but later evidence suggested that this complex might be part of the prevacuolar compartment [15]. These two studies focus on only a single SNARE protein, and lack of evidence presenting the complete complex may explain the inconsistent conclusions. With respect to such studies, neither forward nor reverse genetics approaches may be able to identify the entire SNARE-pin complex because mutations in genes encoding SNAREs in plants have been shown to be embryo or gametophyte lethal [16]. Proteomics can offer a method to study the SNARE-pin protein complexes throughout the endomembrane system. As such, proteomics profiling of the tonoplast has provided a putative composition of a SNARE-pin complex responsible for membrane fusion at the vacuolar membrane. All four types of SNARE proteins (Q-, v-, t-, and R-SNARE), and only one of each type, that are required for the complex formation and membrane fusion are found in the tonoplast [6]. In addition, several proteins that likely interact with the SNARE-pin complex, including GTP-binding proteins, are also identified. These data provide strong evidence that the identified SNARE-pin complex and its associated proteins are indeed localized to the tonoplast.

Barley

Barley is the only crop plant whose vacuole proteome has been investigated at a large scale [7]. In this study, the authors isolate tonoplast from the leaf mesophyll cells, and the proteins are separated by 1D SDS-PAGE. After protein digestion by trypsin, all tryptic peptides are analyzed by LC-MS/MS. As expected, many proteins discovered in the *Arabidopsis* leaf tonoplast are also found in the barley mesophyll tonoplast. The same 11 V-type ATPase subunits are found in both *Arabidopsis* and barley tonoplasts [6, 7], thus, it appears that the two studies are comparable in detection limit. However, overall number of tonoplast proteins is significantly lower in barley than in *Arabidopsis*, 88 versus 219 in total. This may indicate that many proteins present in *Arabidopsis* tonoplast may not exist in barley or that these proteins share low similarity so that identification has been impossible due to the lack of a complete sequence of the barley genome. Another reason for the observed differences may be that the *Arabidopsis* study includes all cell types found in leaves, not just a single cell type of mesophyll cells used in the barley study.

Analyses of all 88 proteins identified from barley tonoplast suggest that majority of these proteins share homology with proteins from *Arabidopsis* [7]. Only six proteins do not have homologues to proteins from *Arabidopsis*, but five of these are hypothetical proteins of unknown function. Forty of the 88 proteins present in barley have not been found in either *Arabidopsis* leaf or suspension cells, suggesting that these proteins are relatively more abundant in barley than in *Arabidopsis*. Among these novel membrane proteins, the CLC-type chloride channel and the sucrose transporter are for the first time discovered in plant vacuoles [7]. Further GFP localization experiments have confirmed that the sucrose transporter is indeed a vacuole-specific protein. Another interesting finding is that differential localization may occur in the same protein family. For example, it is known that the peptide transporter AtPTR1 is localized to PM, but another family member AtPTR2-B and its barley homologue are found in vacuoles of either *Arabidopsis* or barley [6, 7].

Rice

Comparative proteomics has been used to investigate the effects of the plant hormone GA treatment on the rice root tonoplast proteome [11]. In this study, the authors employ the discontinuous sucrose/sorbitol gradient centrifugation to isolate vacuolar membrane fractions. The tonoplast proteins are then separated by 2-DGE and visualized with silver staining. A total of 10 protein spots show increased abundance in the GA-treated roots compared to control. These proteins are identified by either MALDI-TOF-MS or Edman degradation. The identified proteins are FBP aldolase C-1 and V-H1-ATPase, suggesting that the rice root vacuoles are involved in GA signaling. This preliminary study is focused on only GA-induced changes, but the total protein constituents of the rice vacuole proteome are still unknown.

27.3 METHODOLOGY AND STRATEGIES

General Considerations

Isolation of Vacuoles. An important consideration in defining the vacuole proteome is whether the vacuole samples are technically pure. It is crucial to keep other potential contaminants below the detection limit. Practically, purification and isolation of plant vacuoles have been somewhat standardized and routinely used in recent proteomics studies, where the gradient centrifugation is a general method to purify and isolate intact vacuoles from plant tissues or suspension cultures [6–11]. However, different groups still show different levels of protein contaminants from other organelles such as mitochondria and chloroplast, although marker proteins for these organelles are not detected when checked with Western blot analysis or enzymatic assays. For example, in the study by Shimaoka et al. [8] a large number of glycosidases and proteases (mostly soluble proteins) are identified in the *Arabidopsis* tonoplast proteome. This preparation also contains some PM protein such as AtSUC1 that is confirmed to be PM location by a GFP fusion experiment in a later study [7]. It is possible that the finding of these proteins may have been caused by sample handling, such as lack

of thorough washing of the tonoplast fraction. It is important for experimentalists to understand the necessity to prepare vacuoles with the highest possible purity.

LC-MS/MS versus MALDI-TOF-MS. In vacuole proteomics studies, both LC-MS/MS and MALDI-TOF-MS techniques are used. Most investigators have used 1D SDS-PAGE to separate total vacuole or tonoplast proteins followed by trypsin in-gel digestion and then used the LC-MS/MS analyses for protein identifications. In a study by Szponarski et al. [9], tonoplast proteins are resolved by three-step separation: gel filtration, anion exchange chromatography, and 1D SDS-PAGE. A total of 200 CBB stained gel bands are visible after these multi-separation procedures. Finally, the MALDI-TOF PMF is used, and it is able to identify only 70 proteins. On average, the success rate of protein identification using this PMF technique is one protein per three sample bands [9], compared to more than four proteins per band by using the LC-MS/MS technique [6]. Thus, we believe that the LC-MS/MS approach generally has an advantage in sensitivity over the MALDI-TOF method in identifying vacuole proteins.

Importance of High-Resolution MS. MS is the key technology that is essential to all proteomics studies. Success in protein identification and quantification is largely dependent on the types and use of the MS instruments. High-quality MS and MS/MS spectra are essential to the accuracy of peptide sequence assignment. Abundance of cellular proteins, including vacuolar proteins, varies as much as 10^4 in dynamic range [17]. Many low abundance proteins have limited sequence coverage after in-gel trypsin digestion, and very often these proteins may have only a single peptide ion detected or be below the detection limit. Identification of these proteins still represents a significant challenge in today's proteomics technologies.

Historically, MS/MS-based peptide sequencing is first performed with relatively low-resolution instruments such as the ion trap-MS. Because of low resolution, ambiguous assignment of ion charge states and m/z values can occur, which may lead to incorrect protein identification. Many researchers using this type of instrument emphasize that there must be a minimum of two peptide hits for each protein identification in order to avoid FPs. However, this may exclude many low-abundance proteins with only single peptide detected by LC-MS/MS. To rescue these low-abundance proteins from being excluded in proteomics profiling, we believe it is critically important to use high-resolution MS systems such as triple Q-TOF. For high-resolution instruments, charge states and m/z values of peptide ions can be accurately determined. Therefore, protein identification with a single MS/MS spectrum can become reliable as long as the fragmentation patterns provide enough information for daughter ions assignment. In our study, approximately 46.5% of total vacuolar proteins are identified by single-peptide hits including some SNARE proteins [6].

Defining Vacuole Proteome: The LOPIT Strategy

In a study of ER and Golgi proteins, Lilley and co-workers [18] have developed the LOPIT strategy to discriminate the two types of proteins. One major challenge

to organelle proteomics is that organelle cross-contamination often compromises the confidence in assignment of individual proteins to a specific organelle, particularly for those proteins whose cellular localization is not known. This challenge is even more profound in studies of the plant endomembrane system including the vacuole, ER, and Golgi. Different compartments of the endomembrane system have very small difference in physical density that prevents them from complete purification by gradient centrifugation. The LOPIT strategy is shown to be an effective method to tackle this challenge. The principle of the LOPIT strategy is based on quantitative ICAT. For example, the Lilley group [18] uses the heavy ICAT tag to label proteins from the Golgi fraction and the light ICAT tag to label proteins from the ER fraction. The degree of cross-contamination can be estimated by comparing the relative heavy/light ICAT ratio for the ER- and Golgi-specific protein markers. Other proteins whose heavy/light ICAT ratios are close to the ER markers are considered to localize to the ER, or otherwise to the Golgi if the ratios are similar to the Golgi markers. They tested 28 proteins with known or predicted localization; the LOPIT method was able to assign them correctly to Golgi, ER, or mitochondria/plastid contaminants. Presumably, this strategy can also be applied to studies of the vacuole proteome. The tonoplast integral proteins and V-type ATPase subunits can all serve as specific vacuolar markers.

Improving Sequence Coverage of Peptides and Proteins

Due to the wide dynamic range and PTMs of cellular proteins and the limitations in analytical capability with MS-based technologies, currently no single proteomics study can be claimed as a complete characterization of any given proteome. However, we believe that use of multiple and alternative strategies ranging from sample preparation to data acquisition do help increase proteome coverage of plant vacuoles.

Differential Fractionation. As shown in our study, both intact vacuoles and tonoplast fractions are subjected to proteomics profiling. This differential fractionation of the same leaf central vacuoles and subsequent LC–MS/MS analyses has clearly showed advantage in increasing coverage of the vacuole proteome. For the intact vacuoles, a total of 267 proteins are identified while 219 identified in the tonoplast fraction. The two sets of data do not have many proteins in common, but rather have a significant portion of proteins specific to either fraction. When combined together, overall 379 proteins are identified. The increase of proteome coverage by using the two fractions is 42% and 73%, respectively, compared to either intact vacuole or tonoplast alone [6].

Gel-Based Proteomics versus MudPIT. With vacuolar proteins prepared, researchers also have a choice for different separation methods such as gel-based proteomics vs MudPIT. For the intact vacuoles, we have used both methods. The total proteins can be digested in-solution by trypsin and then subjected to MudPIT analyses, or, alternatively, the proteins are separated by 1D SDS-PAGE and digested by trypsin in-gel followed by LC–MS/MS. We have found that the latter method is more effective in profiling the vacuole proteome than the MudPIT in terms of number of proteins identified (267 versus 80). Out of 80 proteins identified by MudPIT,

however, 30 proteins are not found with the gel-based method. Combining the two methods together, the total number of proteins increases to 297, an 11% increase over the number of proteins identified by the gel-based method alone [6].

Differential Scanning. With triple Q–TOF such as the Q–TOF Premier, there is much flexibility with the operation and data acquisition methods. One can repetitively analyze the same sample using different scanning methods. In LC–MS/MS-based proteomics discovery, a general method is to run survey scanning using the data dependent acquisition (DDA). When a complex sample is subjected to DDA, peptide co-elution often occurs and many weak ions can escape from MS/MS analyses. This is particularly common to vacuolar samples derived from 1D gel bands that usually contain multiple proteins. After first survey scanning, we usually run a second survey experiment for the same sample by excluding the peptide ions that have been analyzed in the first DDA. This will allow other co-eluted weak-signal peptide ions to be recognized and sequenced by MS/MS. This second exclusion scanning should be useful in discovering low-abundance proteins and PTMs because their peptide ions signals are generally low. In a study that has systematically evaluated this method, a 25% increase in the number of identified proteins is observed [19]. In agreement with their work, we also consistently improve sequence and protein coverage for the vacuole proteome by this exclusion scanning. An important consideration when using this method is to set an appropriate exclusion window based on the data from the first DDA experiment. However, for simple samples with only few proteins, the improvement may not be significant.

Another way to avoid missing any co-eluted peptides is the so-called “expression” scanning [20]. The instrument scans peptide ions in a single MS mode but alternates between low and high collision energy in the collision cell. At low energy, peptides are detected as precursor ions. Conversely, at high energy they are detected as fragment ions that are corresponding to MS/MS. Sequences of individual peptide precursors can be deduced from the high-energy spectra using an automated program to precisely align the m/z and retention time between a precursor and its fragments. In this case, all co-eluted peptides become fragmented in the high-energy mode without selection or discrimination, which differs from DDA-based MS/MS that is selective scanning for individual peptide precursors. However, the sequence data generated from the “expression” scanning is a mix of many peptide precursors, and highly intelligent software is required to perform the analysis. At present, there is still need for optimization of this approach in order to make it routine in proteomics application.

When comparing different samples for changes in expression of specific proteins, MS/MS scanning may not be required for all samples. When a protein has been definitively identified from one sample, its peptide ions are specifically associated with a particular retention time, m/z ratio, and ion charge states. These values can be compiled as a specific set of so-called accurate mass and retention time (AMRT) tags and can be used for identification of the same protein from other samples in LC–MS experiments without MS/MS [21]. Ion signals in MS mode are generally about 10-fold higher than those in MS/MS mode. So, the advantage of using the AMRT tags over the

MS/MS is that it can significantly increase coverage of very low abundance proteins. We have used this AMRT method to discover many down-regulated vacuolar proteins whose signals may be well below the limit of MS/MS detection. Highly consistent LC performance is very critical to this method. A nonsplitting n-flow LC system is recommended for this approach.

Differential Proteolysis: Trypsin, V8, and Other Endoproteases. Trypsin is an endoprotease that is most commonly used in proteomics studies for protein digestion in-gel or in-solution. Trypsin recognizes only Arg (R) or Lys (K) residues within a protein. Basic proteins may have too many these two residues, and tryptic peptides may be too short for analyses. On the other hand, acidic and highly hydrophobic proteins may contain a very limited number of Arg and Lys residues in their sequences, and tryptic peptides may be too long to be suitable for MS/MS experiments. Under these circumstances, another endoprotease such as V8 can be an alternative to trypsin. It cuts at acidic residues such as Asp and Glu. Moreover, nonspecific endoproteases can also be considered if broader proteolysis is needed. Use of these different endoproteases will generate different sets of proteolytic peptides that are complementary to each other to greatly increase the probability of finding any vacuolar proteins. In addition to differential proteolysis, database searching by selecting semitrypsin instead of trypsin can also uncover tryptic peptides that have undergone in-source fragmentation before entering into the collision cell for MS/MS.

MS-Based Quantitative Approaches

SIL-Assisted MS. SIL is commonly used in coupling of MS in proteomics studies, particularly for quantitative analysis of abundance of cellular proteins between different experimental samples. This SIL-assisted MS approach is widely employed to study differential protein expression related to signal transduction and gene regulation. The general strategy for any SIL-MS method is to induce mass shift by introducing a specific mass tag to cellular proteins and peptides. For quantitative analysis, the proteins in one sample are not tagged, but the proteins in another sample are tagged. Two samples are equally mixed together; MS measurement of the tagged and the nontagged peptide ions can provide a quantitative ratio for the same protein between two samples based on the measured signal intensities.

There are several ways for creating an SIL tag, either *in vitro* or *in vivo* [22]. For example, ^{18}O -labeled water can be used in trypsin digestion to introduce a 2-Da mass tag through the enzymatic reaction to uniformly label all proteolytic peptides at the C-terminus. Similarly, peptides can also be uniformly tagged at their N-terminus through the acetylation reaction, which introduces a 42-Da tag. The Cys-containing peptides can be selectively isolated and characterized using ICAT [23]. The mass-coded abundance tagging (MCAT) can be used to characterize Lys-containing peptides for protein quantification [24]. More recently, isobaric tags for relative and absolute quantitation (iTRAQ) technique has been used to uniformly label the N-terminus of the proteolytic peptides for multiplex quantitative comparison using several mass reporters in MS/MS spectrum [25].

Plant cellular proteins can also be mass-tagged through cellular metabolism in the process of protein translation. Metabolic labeling can be achieved using either ^{15}N universal labeling [26] or SILAC [27]. By using ^{15}N -labeled nitrogen source in growth medium, all plant proteins are mass-tagged during translation. After trypsin digestion, the measured peptide ions will show a certain mass increase. However, the mass shift is not a fixed value because individual peptides may contain different number of nitrogen atoms dependent on their amino acid contents. Quantification using ^{15}N -labeled peptides can be difficult due to this uncertainty. The SILAC strategy tags a particular type of amino acid residue with predictable mass shifts in MS spectra. However, since plants are autotrophic, labeled amino acids in the medium can be metabolized to other compounds after being uptaken into plant cells. Among several amino acids tested, only arginine showed relatively high labeling efficiency with approximately 80% of the labeled arginine incorporated into plant cellular proteins [27]. Overall, the SIL-assisted MS approach has seen limited success in plants and has not been used for vacuolar proteins.

A Label-Free Strategy for Quantitative Proteomics. Since SIL methods need relatively expensive reagents and also require specific sample preparation procedures, an alternative label-free method has been proposed by several groups of researchers including us [28–31]. We believe that this label-free quantitative approach may be of particular value to plants.

A general method for the label-free approach is described as in the diagram (Figure 27.1). For the two samples that need be compared quantitatively, the LC–MS/MS experiment is first performed for both samples to determine protein constituents, and a precursor ions m/z and retention time file is generated for all MS/MS spectra of each identified protein. After the LC–MS/MS experiment, the LC–MS

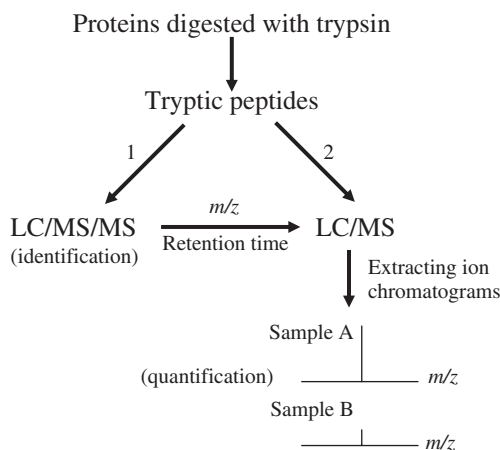


FIGURE 27.1. A schematic diagram to show a general strategy of the label-free quantitative proteomics.

experiment is immediately followed for the quantitative purpose. In the LC–MS spectrum, all peptide ions are individually extracted to retrieve total signal spectral counts using the same principle as for the labeling approach, and these spectral counts are used to quantify the relative difference of proteins in comparison between samples. Unlike the labeling methods with which quantitative analyses are limited to only the tagged peptides, with label-free methods essentially all peptides can be used and the quantitative comparison can be multiplex instead of binary. An example of applying this label-free method in study of an *Arabidopsis* mutant with VPE-depleted vacuoles will be described in the following section.

For the label-free quantitative approach, of most concern is (a) the variability associated with sample preparation and injection and (b) LC–MS experimental consistency. To normalize out these possible variables, spectral counts of several of trypsin's auto-cleavage peptides can be chosen for this purpose if the same amount of trypsin is used for different samples under comparison. For samples containing a large number of proteins, it is expected that only some proteins may be significantly different in abundance while the bulk of the proteins should be similar and can be used for normalization. When estimating change of PTM level of a particular peptide, the non-PTM peptides of the same protein should be chosen for normalization. Finally, researchers are searching for small nonretentive hydrophilic compounds that can be used for normalizing any fluctuations during entire LC–MS experiment. Of course, all of these analyses must require an automated program in order to carry out large-scale multiplex mathematical calculations.

27.4 A CASE-STUDY OF LABEL-FREE QUANTITATIVE PROTEOMICS

Characterization of the *Arabidopsis vpeγ* Mutant

Quantitative proteomics is commonly used to study differential protein expression in human and plants to identify a functional pathway that is triggered by overexpression or suppression of a specific gene [32]. We have employed this approach to determine the biological effects of a T-DNA mutant of a vacuolar protease, VPE γ , in *Arabidopsis* by comparing the wild-type and mutant vacuole proteomes. VPE γ is a protease responsible for processing and degradation of many vacuolar proteins, and its knock-out mutant is expected to result in accumulation of precursors of these proteins in vacuoles. Thus, a quantitative proteomics study may reveal potential VPE γ substrates *in vivo*.

To identify proteins that may be targets in the VPE-dependent pathway, the same amount of total vacuolar proteins from either wild-type or *vpeγ* mutant was subjected to 1D SDS-PAGE separation in two adjacent lanes. The two gel lanes were cut into narrow horizontal slices with corresponding wild-type and *vpeγ* bands processed in a pairwise manner. It is important to cut the bands as precisely as possible for the two lanes so that the same proteins are compared. After routine trypsin in-gel digestion, the tryptic peptides were sequenced by LC–MS/MS and quantified by the label-free LC–MS method. Out of all the proteins identified, our preliminary analyses have found more than a dozen proteins accumulating more than threefold in *vpeγ* compared to

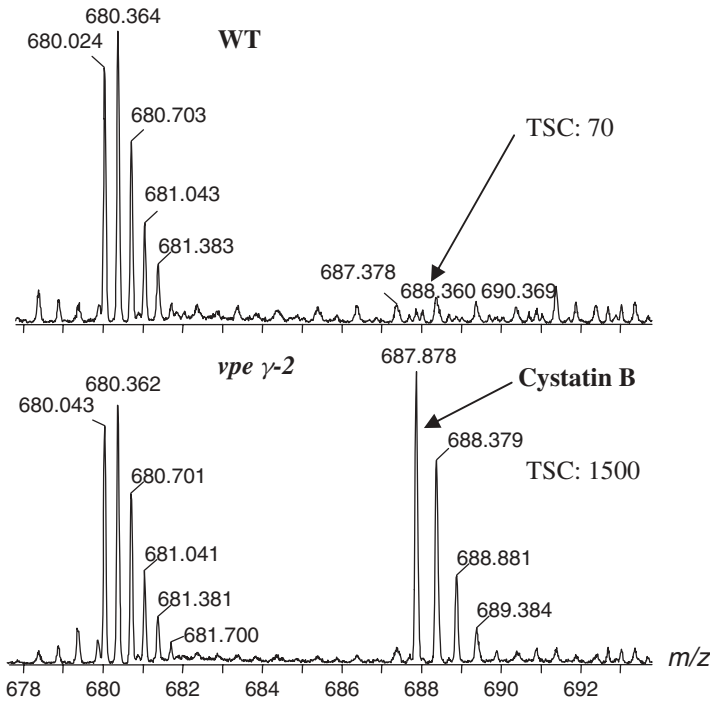


FIGURE 27.2. A tryptic peptide of the cystatin B protein is quantified by the label-free LC/MS spectrum for wild-type vacuole and the *vpe* mutant. TSC counts for m/z 687.88 (2+) that represents cystatin B are approximately 20-fold higher in *vpe* mutant than in wild-type. As an internal control, another co-eluted peptide (m/z 680.04, 3+) shows similar TSC counts (approximately 1500) between the two samples. TSC, total spectral counts. Reprinted with permission [reference 29].

the wild-type vacuole. Among these identified proteins, carboxypeptidase Y (CPY), invertase and seed storage proteins are previously known to accumulate as precursors in the *vpeγ* mutant [33], indicating the applicability of this label-free method in study of the vacuole proteome. Very interestingly, the highest accumulation (>20-fold) in *vpeγ* is the cystatin B protein (Figure 27.2), which is an apoptosis inhibitor in animal cells and may be involved in the VPE-mediated PCD in *Arabidopsis* [29]. Whether cystatin B is a direct substrate of the VPE protease awaits further biochemical studies.

27.5 CONCLUSIONS

Although there are relatively few proteomic studies of the plant vacuole, we have seen that this approach has increased our understanding of this organelle. Assignment of many previously known and unknown proteins to the vacuole expands our view of its internal organization. Proteins identified through proteomic profiling represent a wide range of functions that are well-correlated with biological roles of vacuoles in plant.

Clearly, these studies have demonstrated that proteomics is a powerful approach for characterization of vacuolar proteins on large scales. For example, discovery of many *Arabidopsis* putative endogenous CTPP- and NTPP-containing proteins by proteomics studies may not be possible or at least occur at much slower pace if, otherwise, using the conventional genetics approaches. These proteins can serve as excellent candidates for further functional characterization to understand the molecular mechanisms of the two pathways in *Arabidopsis*. In addition, quantitative proteomics can serve as a screening tool for discovering proteins that may be functionally important to a specific pathway such as the VPE-mediated protein processing in vacuoles. On the technology side, proteomics is much different from genomics, for which cellular DNAs and RNAs can be amplified as much as needed. However, cellular proteins cannot be amplified; therefore, proteomics results are greatly dependent on the detection and performance limit of the entire working system.

27.6 FIVE-YEAR VIEWPOINT

Global systematic investigations of plant vacuole proteomes have only recently emerged and will certainly continue for years. In the near future, research focuses will likely turn to quantitative and comparative analyses of vacuole proteomes and to characterization of PTMs and protein complexes. To understand the cellular mechanisms of trafficking pathways, these proteomics analyses should be conducted in the context of chemical treatments or mutant backgrounds that either enhance or inhibit trafficking activity. Once alterations in vacuole morphology or activity are observed, subsequent proteomics analyses can then follow to identify proteins and PTMs that underline the altered phenotypes. Combined with other systems biology approaches, proteomics will play a major role in dissecting the cellular pathways of protein trafficking and vacuole biogenesis.

We envision that quantitative and comparative proteomics will be crucial to further characterization of plant vacuole proteomes. For example, quantitative comparison of *Arabidopsis* and rice vacuole proteomes may be necessary to discover dicot- or monocot-specific vacuolar proteins that can be an indication of functional specialty between the two angiosperms. The difference in functionality in vacuole between dicot and monocot plants might be well expected. The sucrose transporter protein, SUC2, is found in barley but not in *Arabidopsis*. This fact suggests that the monocot plants might need a more developed sucrose transporter system than would dicot plants [7]. To understand biogenesis and development of vacuoles within plant cells, clearly defining proteome contents of different types of vacuolar compartments such as the prevacuoles, transporting vesicles, lytic vacuoles, and protein storage vacuoles is needed. Again, quantitative comparisons should be carried out to identify any proteins that may mark developmental pathways that lead to biogenesis of a specific vacuolar compartment.

Many so-called “expressed protein” are functionally unknown and they account for about 35% of the entire *Arabidopsis* proteome. This category of proteins also represents a large group in the plant vacuole proteome [6]. Finding novel functions

of these proteins will certainly expand our knowledge of vacuole biology as a whole. One approach to revealing putative functions of these proteins in vacuoles is using comparative and quantitative proteomics. Alteration in expression or accumulation of these proteins in vacuoles in response to developmental dysregulation or biotic/abiotic treatments can be an indication of possible functions of these proteins. This type of investigations should be conducted in context of a specific pathway that can be easily focused. Information obtained from such studies will certainly help uncover the functional mysteries of these “expressed proteins” in the vacuole.

Coupling chemical genomics and functional proteomics should be another area of future focuses. Chemical genomics is also an emerging technology that utilizes small organic compounds to perturb or disrupt normal biologic activities of cellular proteins. These small chemicals are very useful for creating reversible conditional mutants in plants to avoid the synthetic-lethal mutants derived from conventional genetics methods. In particular, using large-scale screening, Raikhel and colleagues have successfully identified compounds such as the “sorting inhibitors” (sortins) that can enhance secretion of a vacuolar protein, CPY [34]. CPY is normally sorted to the plant vacuole via a cleavable NTPP signal. While multiple pathways exist for protein sorting to the plant vacuole, the receptors and specific steps of these pathways remain essentially unknown. The enhanced secretion of CPY suggests that the NTPP pathway has been disrupted by sortins that may target to protein(s) that function in this pathway. Proteomics can thus follow to identify the sortin-targeted proteins after the ligand-affinity purification. Similarly, other ongoing chemical genomics screens targeted to vacuole morphology and functions will uncover a series of conditional mutants with altered phenotypes [35]. Identifications of these chemical-targeted proteins responsible for the phenotypes will also be dependent on proteomics approach. Therefore, combination of high-throughput chemical genomics and proteomics will provide a systems platform for discovering specific proteins and genes that function in plant vacuoles in a pathway-specific manner.

ACKNOWLEDGMENTS

The authors thank Dr. Clay Carter for his critical reading of the manuscript. Research in the authors’ laboratory was supported by a grant from the U.S. Department of Energy, Division of Energy Biosciences (DE-FG02-02ER15295 to N.R.).

REFERENCES

1. Rojo E, et al. 2001. *Dev Cell* **1**:303–310.
2. Surpin M and Raikhel, N. 2004. *Nat Rev Mol Cell Biol* **5**:100–109.
3. Vitale A and Raikhel NV. 1999. *Trends Plant Sci* **4**:149–155.
4. De DN. 2000. *Plant Cell Vacuoles*, CSIRO Publishing, Collingwood, Australia.
5. Maeshima M. 2001. *Annu Rev Plant Physiol Plant Mol Biol* **52**:469–497.
6. Carter C, et al. 2004. *Plant Cell* **16**:3285–3303.

7. Endler A, et al. 2006. *Plant Physiol* **141**:196–207.
8. Shimaoka T, et al. 2004. *Plant Cell Physiol* **45**:672–683.
9. Szponarski W, et al. 2004. *Proteomics* **4**:397–406.
10. Sazuka T, et al. 2004. *DNA Res* **11**:101–113.
11. Konishi H, Maeshima M, and Komatsu S. 2005. *J Proteome Res* **4**:1775–1780.
12. Rojo E, et al. 2002. *Plant Cell* **14**:969–977.
13. Ahmed SU, et al. 2000. *J Cell Biol* **149**:1335–1344.
14. Sato MH, et al. 1997. *J Biol Chem* **272**:24530–24535.
15. Sanderfoot AA, et al. 1999. *Plant Physiol* **121**:929–938.
16. Surpin M, et al. 2003. *Plant Cell* **15**:2885–2899.
17. Baginsky S and Gruissem W. 2004. *J Exp Bot* **55**:1213–1220.
18. Dunkley TP, et al. 2004. *Mol Cell Proteomics* **3**:1128–1134.
19. Chen HS, et al. 2005. *Anal Chem* **77**:7816–7825.
20. Silva JC, et al. 2006. *Mol Cell Proteomics* **5**:144–156.
21. Conrads TP, et al. 2000. *Anal Chem* **72**:3349–3354.
22. Pan S, et al. 2003. *Anal Chem* **75**:1316–1324.
23. Gygi SP, et al. 1999. *Nat Biotechnol* **17**:994–999.
24. Cagney G and Emili A. 2002. *Nat Biotechnol* **20**:163–170.
25. Ross PL, et al. 2004. *Mol Cell Proteomics* **3**:1154–1169.
26. Ippel JH, et al. 2004. *Proteomics* **4**:226–234.
27. Gruhler A, et al. 2005. *Mol Cell Proteomics* **4**:1697–1709.
28. Chelius D, et al. 2003. *Anal Chem* **75**:6658–6665.
29. Rojo E, et al. 2004. *Curr Biol* **14**:1897–1906.
30. Fang R, et al. 2006. *Mol Cell Proteomics* **5**:714–725.
31. Old WM, et al. 2005. *Mol Cell Proteomics* **4**:1487–1502.
32. Gu S, et al. 2004. *Mol Cell Proteomics* **3**:998–1008.
33. Rojo E, et al. 2003. *Proc Natl Acad Sci USA* **100**:7389–7394.
34. Zouhar J, Hicks GR, and Raikhel NV. 2004. *Proc Natl Acad Sci USA* **101**:9497–9501.
35. Surpin M, et al. 2005. *Proc Natl Acad Sci USA* **102**:4902–4907.

OIL BODIES

Pascale Jolivet, Luc Negroni, Sabine d'Andréa, and
Thierry Chardot

28.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Cells store lipids in specialized structures whose name varies depending on considered species. In plants they are called oil bodies (OB), or oleosomes. They consist of a monolayer of phospholipids in which different proteins are embedded [1], surrounding a core of neutral lipids. For a long time they have been considered as balls of fat [2]. Depending on the species, plant seed OB may contain highly variable number of proteins. Recent proteomics studies that will be detailed in this chapter showed that the protein complement of plant seed OB contained from three in castor bean [3] to at least 24 different proteins in *B. napus* [4]. Plant OB proteins fall within two major categories: structural proteins and enzymes, with the function of other minor proteins still being unknown.

Oleosins are the most abundant structural proteins located within plant OB, and their presence is not restricted to seeds resistant to desiccation [5]. They exhibit a conserved architecture among species. One of their functions may be to stabilize OB during seed imbibition prior to germination [6]. In addition to their structural function,

oleosins may also serve as recognition signals on the surface of plant OB for the binding of newly synthesized lipase during germination [7]. In *Arabidopsis*, 16 different oleosins, with molecular mass between 11 and 53 kDa, are encoded by a multigene family [8]. Their mRNAs have been found in seeds, tapetum, and microspore, which lead to the classification of these proteins in four subfamilies corresponding to the location of their mRNAs [8]. Oleosins have also been located in root tips [9]. Interest for oleosins has considerably risen upon their identification outside of seeds. Roberts et al. [10] have sorted oleosins into gametophyte (pollen)-specific and sporophyte (seed)-specific members, with the two gene families either distantly related or of independent origin. They have suggested an analogy between pollen and seeds, both being a dehydrated propagule carrying amounts of stored lipids with a requirement for rapid mobilization of these reserves upon germination. Both pollen and seeds have acquired oleosins as a mean of stabilizing storage material and facilitating its mobilization. One oleosin reported in pollen coat proteome is required for rapid initiation of *Arabidopsis* pollinization [11, 12]; and its C-terminus, highly divergent between homologues and orthologues, makes it a good candidate for species-specific signal.

Another protein, named caleosin, is also found on the surface of OB. It was initially discovered associated with cell-membrane fractions in developing and germinating rice caryopses, in response to ABA or to osmotic stress [13]. The first caleosin described in OB was from sesame seeds [14]. In the case of *Arabidopsis*, caleosin belongs to a multigene family [9, 15]. AtClo1 expression is specific to embryo and mature seeds, and its promoter is active in root tips, where the protein was also immunodetected [9]. Expression of this protein, capable to bind calcium was induced by drought and also by high salinity [16].

The list of enzymes found within plant OB is growing with time. To date, all characterized enzyme activities found associated with OBs use lipids as substrates. The presence of triacylglycerol hydrolase initiating storage oil breakdown in germinating seeds and of enzymes belonging to the short-chain dehydrogenase/reductase family exhibiting steroid dehydrogenase activity will be described later. Some OB proteins are also described by authors under the generic term of “contaminants.” However, their presence may reflect on one hand the vicinity of OB with different cellular compartments and on the other hand the impossibility to obtain pure organelle [4, 17].

28.2 METHODOLOGY AND STRATEGY

Proteome Identification Requires High-Purity OB

To investigate the properties and organization of OB, a method of preparation capable to yield high-purity organelles is necessary. A simple and rapid procedure for the purification of seed OB was first reported by Murphy and Cummins [18]. OB were collected by two floatation steps following centrifugation of tissue homogenates on sucrose gradient. Some OB were difficult to purify to homogeneity using this technique and were notably contaminated by storage proteins. A protocol was developed by

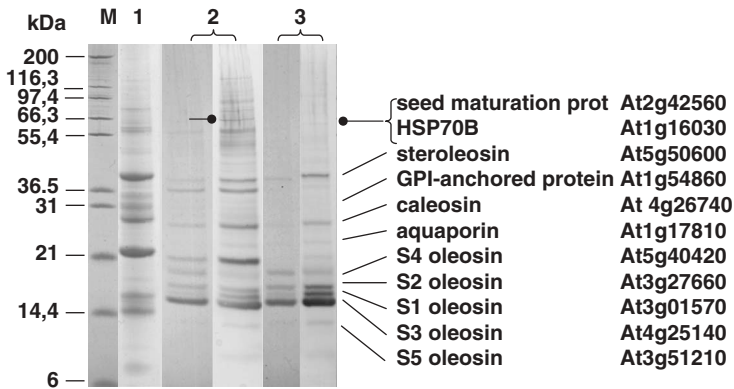


FIGURE 28.1. OB-associated proteins. Proteins from *A. thaliana* seed crude extract (lane 1) and from OB, CF3, or CF6 fraction [21] (lanes 2 and 3) have been separated on PAGE. Protein bands have been digested and analyzed using LC–MS/MS. For lanes 2 and 3, right bands have been silver-stained.

Millichip et al. [19] based upon the stringent washing of the OB fraction in 9 M urea. Finally, OB were purified with six floatation steps including detergent washing (0.1% Tween 20), ionic elution (2 M NaCl), treatment with a chaotropic agent (9 M urea), and hexane washing [20]. The high purity of OB obtained by this new method provides a better source for the investigation on the organization of the organelle. Non-specifically associated proteins are removed, and the proteins identified by proteomic investigation are embedded integral proteins.

The integrity of OB was examined by light and EM [1, 18], and hexane extraction [20] and the removal of contaminating proteins were checked by SDS-PAGE (Figure 28.1) [19, 21]. HSP70 and seed maturation proteins were noticeable contaminant of low purified OB (Jolivet, unpublished results). Tzen et al. [1] reported that OB isolated from different mature seeds had average diameters within a narrow range (0.6–2.0 μm) as measured from EM. These authors showed that seed OB from diverse species were very similar in structure and possessed a negatively charged surface at neutral pH with pI of 5.7 to 6.6. OB contain about 92–98% neutral lipids (especially triacylglycerols and subsidiary diacylglycerols and free fatty acids), 1–4% phospholipids (phosphatidylcholine is major, and phosphatidylserine, phosphatidylethanolamine and phosphatidylinositol are minor), and 1–4% proteins [22].

Separation and Identification of Highly Hydrophobic and/or Alkaline Proteins. The classical separation techniques, using electrophoresis or LC in aqueous buffers, have given poor results with hydrophobic proteins in comparison with hydrophilic proteins. But MS, SDS-PAGE, and multidimensional chromatography have been the tools for significant advances. 2-DGE is by far the most powerful technique for protein separation. Unfortunately, hydrophobic proteins are usually insoluble in the buffer used for the IEF step, even if some improvements were obtained

with the use of non-ionic detergents and chaotropic agents [23, 24]. Regarding the proteomics of hydrophobic complexes, various membrane complexes or fractions were analyzed successfully by BN-PAGE [25–27].

The major MS approach is to analyze peptides instead of intact proteins. This is quite convenient with hydrophobic proteins because the enzymatic digestion usually generates hydrophilic peptides from the cytosolic domains, loops, or extremities. Mono- or multidimensional chromatography resolves the complexity of peptide extracts before identification with MS/MS. Trypsin is generally used with modified protocols enhancing the digestion of hydrophobic proteins [28]. Other proteases, such as proteinase K, may be an alternative when no Lys or Arg are properly distributed in the protein [29, 30]. Either hydrophobic proteins are separated using SDS-PAGE and individual bands are digested and analyzed with LC–MS/MS [21, 31–34] or protease digests from protein fractions are directly analyzed with mono- or multidimensional chromatography [35, 36].

Proteome of Seed OB

Exhaustive protein compositions from OB from castor beans (*R. communis*) [3], *A. thaliana* [21], and *B. napus* [4, 37] were recently determined. Proteins from purified OB were often precipitated overnight in cold 100% acetone (–20°C) and subsequently solubilized in electrophoresis or IEF sample medium. Proteins were separated by SDS-PAGE or 2-DGE. Analyses of trypsin peptides obtained after in-gel digestion were performed using MALDI or LC–MS/MS devices. For protein identification, acquired mass spectra were searched against the NCBI nonredundant database or against more specific ESTs database. Major proteins were identified as oleosins, but some other minor proteins were also recognized.

Structural Proteins. Oleosins are alkaline proteins of low molecular mass (15–26 kDa). They all possess three structural domains [7]. Each oleosin molecule has a highly conserved central hydrophobic domain of 68–74 amino acid residues, which is the longest hydrophobic sequence found in any organism. This domain was proposed to be either an anti-parallel α -helix penetrating into the TG matrix or a long stretch of β -strand structure running underneath the surface of OB. The loop in the anti-parallel β -strands has a “proline knot” consisting of three proline and one serine residues. An amphipathic domain of 40–60 amino acids, the secondary structure of which is unknown, is present at the N-terminus and likely associated with the OB surface. An amphipathic α -helical domain is situated at the C-terminus. The positively charged residues face the negatively charged lipids on the boundary phospholipids layer, whereas the negatively charged residues are exposed to the surface. Together, the N-terminal, the central, and the α -helical portions anchor the oleosin stably on the surface of the OB. Two distinct oleosin isoforms were first described in maize (16–18 kDa [38]), soybean (18–24 kDa [38]), sesame (15–17 kDa [39]), and castor bean (14–16 kDa [3]), only one oleosin was recovered in almond OB [40], and numerous cDNA and genomic sequences encoding for these proteins have been reported from various species. Sixteen oleosin genes have been characterized in the

A. thaliana genome, with five genes specifically expressed in maturing seeds [8]; five genes of the coffee oleosin family have been very recently described, with *CcOLE-1* and *CcOLE-2* being more highly expressed [41]. Regarding the proteomic analysis of highly purified seed OB, two oleosins were identified in castor bean OB [3], five in *A. thaliana* [21], and three to eight in *B. napus*, depending on database used for protein identification [4, 37].

Caleosin gene sequences have been found in a wide range of plants and oil-accumulating fungi [9] composing a multigene family in the case of *A. thaliana* [9, 42], *B. napus* [43], or barley [44]. Caleosins contain an N-terminal region with a single calcium-binding EF-hand, a central hydrophobic region with a potential membrane anchor, and a C-terminal region with conserved protein kinase phosphorylation sites [13]. The presence of the hydrophobic region suggested an oleosin-like association with OB. Immunological detection and proteomic analyses suggested that caleosin was unique in seed OB from sesame, sunflower, soybean, peanut [45], rapeseed [4, 45], and *A. thaliana* [21]. Caleosins could have a potential role in lipid trafficking, membrane expansion, and OB formation during seed development, as hypothesized by Liu et al. [44]. Recent works proved the involvement of AtClo1 in the degradation of storage lipid in OB [46]. According to Hanano et al. [47], AtClo1 could be involved in the biosynthesis of phytooxylipins and capable of catalyzing hydroperoxide-dependent mono-oxygenation reactions that are characteristic of peroxxygenases.

Enzymes. Besides abundant oleosins and caleosin, proteins named steroleosins for their sterol-binding capacity have been more recently described in OB of sesame, *A. thaliana*, and rapeseed [4, 21, 37, 48, 49]. These proteins possess a hydrophobic anchoring segment followed by a soluble domain homologous to sterol-binding dehydrogenases/reductases with an active site comprised between a NADPH-binding and a sterol-binding subdomain [48]. Sterol-coupling dehydrogenase activity was demonstrated in purified OB from sesame or *A. thaliana* seeds as well as in the overexpressed soluble domain of steroleosin [48, 50]. According to enzymatic activity assays and sequence homology to hydroxysteroid dehydrogenases, steroleosins are classified as 11 β - and 17 β -hydroxysteroid dehydrogenases and belong to a very large group of NAD(P)-dependent oxidoreductases, the short-chain dehydrogenase/reductase family. Moreover, *A. thaliana* steroleosin was described to catalyze NADPH-dependent 17 β -ketosteroid reduction [50]. The question of the physiological substrates and functions of steroleosins in plants remain to be elucidated even if Lin and co-workers [48, 49] hypothesized their involvement in plant signal transduction regulated by various sterols. In a recent article, Katavic et al. [4] have reported in *B. napus* the presence of (a) three isoforms of 11 β -hydroxysteroid dehydrogenase, (b) a new short-chain dehydrogenase/reductase, and (c) a myrosinase-like protein showing similarity with the lipase hydrolase family of enzymes with GDSL motif.

Lipase activities have been early characterized in seeds from rape, mustard, or corn and were reported to be associated to OB [51–53]. Enzyme activities are only detectable upon germination and increase concomitantly with TG breakdown. However, until now, only few genes encoding these enzymes have been cloned and characterized [3, 54, 55]. Eastmond [3] recently reported a relatively amphipathic

nature of OB-associated lipase from castor bean. He suggested that the N-terminus region of protein, devoid of “proline knot” and less hydrophobic than oleosins central region, might play a role in anchorage to OB.

Valencia-Turcotte and Rodriguez-Sotres [56] demonstrated the presence of a diacyl glycerol acyl transferase activity in purified maize OB. Finally, an almond 9-hydroperoxide lyase associated with OB has recently been reported [57] with the targeting of the protein verified through GFP fusions transiently expressed in tobacco protoplasts.

Minor Proteins. Other proteins such as an aquaporin and a putative GPI-anchored protein [21] have been detected in purified OB from *A. thaliana* upon silver staining, which proves that these proteins existed only in very low abundance. The presence and function of aquaporin remain to be confirmed: this protein is proposed to isolate OB from water or, on the contrary, to facilitate the intake of water and uptake of glycerol during lipolysis and germination. GPI-anchored proteins are often hardly detected in classical proteomics analyses. Some are not recovered from 2D gels and must be analyzed by 1D SDS-PAGE. Moreover, trypsin digestion of these proteins yields few or no peptides for LC-MS/MS analysis due to their high level of glycosylation. Some GPI-anchored proteins could be involved in lipid transfer or associated with membrane lipid microdomains.

Investigation of OB Structure and Function

PTMs of OB Proteins. N-terminal sequence of oleosins from rape, sunflower, or almond [19, 40, 58] was found to be blocked before the existence of any structural modifications has been reported in the case of sesame [59]. These authors have found that N-termini of caleosin and 17-kDa oleosin were acetylated on alanine residue after the removal of the first methionine. In addition, deamidation of a glutamine residue was observed in the case of oleosin. On the other hand, steroleosin was not modified. We have also recently observed a probable acetylation of N-terminal alanine residue for one rape oleosin very homologous to *A. thaliana* S1 oleosin. In the case of *A. thaliana* steroleosin, the N-terminal sequence was verified after proteinase K digestion and we did not find any significant modifications like deamidation of glutamine or asparagine residues, acetylation, phosphorylation, or methylation in the observed peptides covering 82% of the protein sequence [50].

Targeting of OB Proteins. The OB arise from specific microdomains of the ER membrane that contain the full complement of triacylglycerol-biosynthesis enzymes. Targeting of oleosin and caleosin to seed OB has been investigated either *in vivo* or *in vitro*. Van Rooijen and Moloney [60] used different constructs of *A. thaliana* oleosins lacking one of three structural domains of natural oleosins and introduced into *B. napus*. Complete construct underwent correct subcellular targeting. Removal of the C-terminal domain of oleosin had no effect on targeting. In contrast, loss of the N-terminus or hydrophobic domain resulted in impaired targeting of the protein to the OB. *S. cerevisiae* was used as a heterologous eukaryotic system to study the

intracellular targeting of oleosins [61, 62], and it appeared that neither the N- nor the C-terminal hydrophilic domains are important for targeting and/or membrane insertion. Chen and Tzen [63] have observed the incorporation of oleosins and caleosin to artificial oil emulsions. As for oleosins, the subdomain comprising the “proline knot” motif is crucial for correct caleosin incorporation into OB.

Different authors have been taking advantage of the fact that oleosins could be expressed as fusion proteins correctly targeted to plant OB. Van Rooijen and Moloney [60] were pioneering in demonstrating correct targeting of oleosin-GUS fusion protein in plant OB. Then, Parmenter et al. [64] were able to produce a biologically active hirudin in *B. napus* seeds. The group of Tzen has been demonstrating the incorporation of oleosin fusion proteins into artificial OB and the use of such system to purify proteins expressed in *E. coli* [65, 66].

Steroleosin possesses a noncleavable N-terminal sequence putatively responsible for ER targeting via the signal-recognition-particle (SRP)-dependent pathway and anchors to OB mainly by this N-terminal hydrophobic domain. It is possible that steroleosin and caleosin/oleosin may be anchored to maturing OB through various locations of ER membrane via different auxiliary targeting machinery [59].

Structural Proteomics for Investigation of Oleosins Insertion within OB.

The secondary structure of oleosins purified from sunflower has been investigated by determining CD and FTIR spectra [19], and the spectra obtained were typical of proteins rich in α -helix. The validity of these results has been questioned based on the harsh conditions used to purify the oleosins (treatment with urea), and the removal of the OB protein environment. These conditions could result in destabilization of the native secondary structure. Lacey et al. [67] have therefore used FTIR spectroscopy to determine the secondary structures of oleosins present in intact OB preparations or proteinase-treated ones. They have observed that after proteinase treatment of OB, the central hydrophobic domain of the oleosins was protected from digestion and was likely to be buried within the OB. More, these protected fragments contained almost only α -helical structure. The hydrophobic domain is probably composed of two α -helices with the “proline knot” forming a 180° turn [67]. Native oleosins are proposed to move laterally within the ER membrane until they flow into a region of TG enclosed between the ER membrane leaflets and are subjected to a transition from a constrained to a relaxed state [68]. Using high pH and the proteinase K method described by Wu et al. [30], we have digested *A. thaliana*-purified OB and have obtained peptides belonging to the hydrophobic region of oleosins, verifying their embedded localization (Jolivet, unpublished results).

28.3 CONCLUSIONS

Proteomics techniques and approaches have permitted the identification of a growing number of proteins associated with plant seed OB. Structural proteins involved in OB stabilization, mainly oleosin, the most abundant found in all species studied to date and to a lesser extent caleosin. The number of their isoforms present

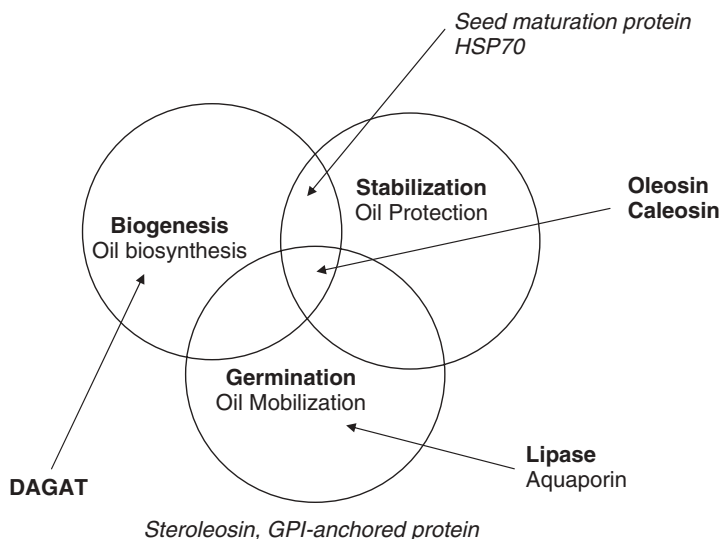


FIGURE 28.2. Function of known OB proteins through OB life cycle. Proteins have been assigned to the different steps of seed OB life cycle. Boldface: proved. Italicized: hypothesis.

within OB varies strongly between species. Enzymes represent a minor part of OB proteins, and this makes them very difficult to identify. They often use lipids as substrates, as illustrated by the recent discoveries of OB-associated triacylglycerol lipase, sterol dehydrogenase, DAGAT activity, and to a lesser extent myrosinase-like lipase hydrolase. Less is known about a third category of proteins, comprising transporters, proteins involved in seed maturation, and protein folding (HSP70 chaperone and seed maturation protein, Jolivet, unpublished). Taken together, these proteomics data, obtained to date mostly on OB from mature seeds, combined with appropriate biochemical and molecular biology investigations, naturally led to the picture of OB as an organelle, with its own dynamic (Figure 28.2).

28.4 FIVE-YEAR VIEWPOINT

Toward a Better Understanding of the Role of OB Proteins from Reserve Accumulation to their Mobilization

The role of OB proteins in developing and germinating seeds is being investigated. The first studies have been carried out using either oleaginous plant strains having diverse oil contents [69] or seed lines in which the major oleosin is ablated or severely attenuated [70]. The first results indicated that the size and shape of the OB are determined by the ratio of oils to oleosins synthesized during seed maturation. Oleosin suppression resulted in an aberrant phenotype of embryo cells that contain unusually large OB. Moreover, alteration in accumulation of lipids and proteins caused delay

in germination. However, a precise pattern of proteins associated with OB during biogenesis is still missing. During seed germination, oleosins degradation could be followed through proteomics at the early stages of OB mobilization. Within the first 4 days of seedling growth, oleosin degradation and accumulation of protease-protected oleosin degradation products could be observed [71, 72]. OB were then more susceptible to lipase. During germination, caleosin could also play a role in the degradation of storage lipid, in particular, by promoting specific interaction of OB with vacuoles [46]. Genome sequencing projects led to the task of assigning molecular and cellular functions to many novel enzymes for which knowledge of active site structure and mechanism is limited. A chemical strategy that aims to synthesize and apply small molecule probes has been reported to analyze enzyme functions in samples of high biocomplexity [73, 74]. Functional proteomics should be able to monitor dynamics in protein functions in complex proteomes.

Oleosins and Allergy

Plant seeds and tree nuts are responsible for the vast majority of food allergies. The prevalence of peanut and tree nut allergy has doubled among the children in the United States from 1997 to 2002, reaching 1.2% [75]. Considerable research has been conducted to identify and characterize nut and seed allergens [76]. Most of these allergens belong to the cupin and the prolamin superfamilies, but a new family, oleosin, has recently been identified [77]. The first suspicion of oleosin involvement in allergy was reported more than 10 years ago [78], but awaited 2002 for formal evidence. Pons et al. [79] showed specific IgE-binding with peanut oleosin in three of 14 sera of patients with allergic reaction to peanuts. Interestingly, the IgE binding was weak against oleosin monomers but stronger with proteins of HMW, presumed to be oleosin oligomers. The authors also argued for a possible cross-reactivity between oleosins from various seeds. More recently, oleosins were identified as allergens in two other seeds: sesame seed [80] and hazelnut [81]. The IgEs of all studied patients with sesame allergy bound monomeric oleosins from purified sesame OB, but some of these patients had negative skin prick tests to aqueous sesame extracts. One hypothesis is that the highly hydrophobic oleosins are not present in aqueous seed extracts, due to their weak solubility in saline solutions. Moreover, several findings suggested that oleosins could be in peanut, soybean, and sesame oils and could be involved in allergic reactions to these oils [78, 79]. This raises the question of the quality of seed extracts used for diagnosis. This strongly supports the need of improving extraction and solubilization methods of hydrophobic proteins, together with their proteomics analysis, especially because some of these proteins (i.e., oleosin) recently appeared to be allergens.

REFERENCES

1. Tzen J, et al. 1993. *Plant Physiol* **101**:267–276.
2. Beckman M. 2006. *Science* **311**:1232–1234.

3. Eastmond PJ. 2004. *J Biol Chem* **279**:45540–45545.
4. Katavic V, et al. 2006. *Proteomics* **6**:4586–4598.
5. Guilloteau M, et al. 2003. *Plant Science* **164**:597–606.
6. Leprince O, et al. 1997. *Planta* **204**:109–119.
7. Huang AHC. 1996. *Plant Physiol* **110**:1055–1061.
8. Kim HU, et al. 2002. *J Biol Chem* **277**:22677–22684.
9. Næsted H, et al. 2000. *Plant Mol Biol* **44**:463–476.
10. Roberts MR, et al. 1993. *Plant J* **3**:629–636.
11. Mayfield JA and Preuss D. 2000. *Nat Cell Biol* **2**:128–130.
12. Mayfield JA, et al. 2001. *Science* **292**:2482–2485.
13. Frandsen G, et al. 1996. *J Biol Chem* **271**:343–348.
14. Chen ECF, et al. 1998. *Plant Cell Physiol* **39**:935–941.
15. Day IS, et al. 2002. *Genome Biol* **3**:Research0056.
16. Takahashi S, et al. 2000. *Plant Cell Physiol* **4**:898–903.
17. Kalinski A, et al. 1992. *J Biol Chem* **267**:12068–12076.
18. Murphy DJ and Cummins I. 1989. *Phytochemistry* **28**:2063–2069.
19. Millichip M, et al. 1996. *Biochem J* **314**:333–337.
20. Tzen JT, et al. 1997. *J Biochem (Tokyo)* **121**:762–768.
21. Jolivet P, et al. 2004. *Plant Physiol Biochem* **42**:501–509.
22. Huang AHC. 1992. *Plant Mol Biol* **43**:177–200.
23. Babu GJ, et al. 2004. *Anal Biochem* **325**:121–125.
24. Rabilloud T, et al. 2006. *Methods Mol Biol* **355**:111–120.
25. Fandino AS, et al. 2005. *J Mass Spectrom* **40**:1223–1231.
26. Granvogl B, Reisinger V and Eichacker LA. 2006. *Proteomics* **6**:3681–3695.
27. Pyndiah S, et al. 2006. *Mol Cell Proteomics* **6**:193–206.
28. Blonder J, et al. 2004. *Proteomics* **4**:31–45.
29. Distler AM, et al. 2006. *Anal Biochem* **356**:18–29.
30. Wu CC, et al. 2003. *Nat Biotechnol* **21**:532–538.
31. Borner GH, et al. 2005. *Plant Physiol* **137**:104–116.
32. Athenstaedt K, et al. 2006. *Proteomics* **6**:1450–1459.
33. Marmagne A, et al. 2004. *Mol Cell Proteomics* **3**:675–691.
34. Valot B, et al. 2006. *Proteomics* **6**:S145–S155.
35. Carter C, et al. 2004. *Plant Cell* **16**:3285–3303.
36. Fischer F, et al. 2006. *Mol Cell Proteomics* **5**:444–453.
37. Jolivet P, et al. 2006. *OCL* **13**:426–430.
38. Tzen JTC, et al. 1990. *Plant Physiol* **94**:1282–1289.
39. Peng C-C and Tzen JTC. 1998. *Plant Cell Physiol* **39**:35–42.
40. Beisson F, et al. 2001. *Biochim Biophys Acta* **1531**:47–58.
41. Simkin AJ, et al. 2006. *J Plant Physiol* **163**:691–708.
42. Nuccio ML and Thomas TL. 1999. *Plant Mol Biol* **39**:1153–1163.
43. Hernandez-Pinzon I, Patel K, and Murphy DJ. 2001. *Plant Physiol Biochem* **39**:615–622.
44. Liu H, et al. 2005. *Planta* **221**:513–522.

45. Chen JC, Tsai CC, and Tzen JT. 1999. *Plant Cell Physiol* **40**:1079–1086.
46. Poxleitner M, et al. 2006. *Plant J* **47**:917–933.
47. Hanano A, et al. 2006. *J Biol Chem* **281**:33140–33151.
48. Lin LJ, et al. 2002. *Plant Physiol* **128**:1200–1211.
49. Lin LJ and Tzen JT. 2004. *Plant Physiol Biochem* **42**:601–608.
50. d'Andréa S, et al. 2007. *Biochimie* **89**:222–229.
51. Lin Y-H and Huang AHC. 1983. *Archiv Biochem Biophys* **225**:360–369.
52. Lin Y-H and Huang AHC. 1984. *Plant Physiol* **76**:719–722.
53. Hills MJ and Murphy DJ. 1988. *Biochem J* **249**:687–693.
54. Eastmond PJ. 2006. *Plant Cell* **18**:665–675.
55. El-Kouhen K, et al. 2005. *FEBS Lett* **579**:6067–6073.
56. Valencia-Turcotte L and Rodriguez-Sotres R. 2001. *Biochim Biophys Acta* **1534**:14–26.
57. Mita G, et al. 2005. *J Exp Bot* **56**:2321–2333.
58. Murphy DJ, et al. 1991. *Biochim Biophys Acta* **1088**:86–94.
59. Lin LJ, et al. 2005. *Plant Physiol Biochem* **43**:770–776.
60. Van Rooijen GJ and Moloney MM. 1995. *Biotechnology (New York)* **13**:72–77.
61. Ting JT, et al. 1997. *J Biol Chem* **272**:3699–3706.
62. Beaudoin F and Napier JA. 2002. *Planta* **215**:293–313.
63. Chen JC and Tzen JT. 2001. *Plant Cell Physiol* **42**:1245–1252.
64. Parmenter DL, et al. 1995. *Plant Mol Biol* **29**:1167–1180.
65. Chiang CJ, et al. 2006. *Enzyme Microb Technol* **39**:1159–1165.
66. Chiang CJ, et al. 2007. *Protein Expr Purif* **52**:14–18.
67. Lacey DJ, et al. 1998. *Biochem J* **334**:469–477.
68. Abell BM, et al. 2004. *Plant J* **37**:461–470.
69. Ting JT, et al. 1996. *Planta* **199**:158–165.
70. Siloto RM, et al. 2006. *Plant Cell* **18**:1961–1974.
71. Sadeghipour HR and Bhatla SC. 2002. *Plant Cell Physiol* **43**:1117–1126.
72. Sadeghipour HR and Bhatla SC. 2003. *Plant Physiol Biochem* **41**:309–316.
73. Adam GC, et al. 2004. *J Am Chem Soc* **126**:1363–1368.
74. Birner-Gruenberger R, et al. 2005. *Mol Cell Proteomics* **4**:1710–1717.
75. Sicherer SH, Munoz-Furlong A, and Sampson HA. 2003. *J Allergy Clin Immunol* **112**:1203–1207.
76. Sathe SK, Kshirsagar HH, and Roux KH. 2005. *J Food Sci* **70**:R93–120.
77. Breiteneder H and Radauer C. 2004. *J Allergy Clin Immunol* **113**:821–830.
78. Guéant JL, et al. 1995. *Rev fr Allergol* **35**:312–319.
79. Pons L, et al. 2002. *Allergy* **57**:88–93.
80. Leduc V, et al. 2006. *Allergy* **61**:349–356.
81. Akkerdaas JH, et al. 2006. *Mol Nutr Food Res* **50**:18–23.

PART V

MODIFICATION PROTEOMICS

PHOSPHOPROTEINS: WHERE ARE WE TODAY?

Florian Wolschin and Wolfram Weckwerth

29.1 INTRODUCTION

Since the breakthrough discoveries of reversible protein phosphorylation by Fischer and Krebs, research on protein phosphorylation in general and on the dynamic interplay between kinases and phosphatases has been gaining more and more interest. This research has extended from initial studies describing single phosphorylation/dephosphorylation events to the complex regulations involving multisite phosphorylation in signaling cascades, which are accepted as one of the major regulatory principles of life. Because of the importance of protein phosphorylation, new methods for the analysis of this remarkable modification have been developed continuously. With the ability to enrich and sensitively detect and quantify phosphorylated proteins and peptides, as well as to pinpoint phosphorylation sites using MS, the methods for a more thorough investigation of protein phosphorylation and their dynamics in general are now available. Many vital processes such as cell trafficking, cell signaling, cell cycle control, enzyme activation/deactivation, cation binding, degradation, apoptosis/defense, and others were found to be directly connected to protein phosphorylation [1, 2]. Protein phosphorylation can occur on several different amino acids including

the well-studied Ser, Thr, Tyr, His, and Asp phosphorylation. Less commonly investigated are phosphorylations located on Lys, glutamic acid, aspartic acid, and Cys [3, 4]. Compared to single-site phosphorylation, multisite phosphorylation enables the cell to develop and maintain complex regulatory pathways contributing to the different levels of hierarchical organization. Most of the work has so far focused on mammals and bacteria, and investigation of protein phosphorylation in plants has been considerably less thorough.

There is a wide variety of techniques available to study protein phosphorylation; and most, if not all, of them also apply to the analysis of multisite phosphorylation. The following paragraphs will deal with the most prominent methods. Because protein phosphorylation research is far more widespread in animal than in plant science, some of the given examples are not directly related to plants. However, all of these methods can be applied to plant tissue, and most of them will probably find their way into plant research.

29.2 PHOSPHOPROTEIN AND PHOSPHOPEPTIDE ENRICHMENT

Due to the low abundance of phosphorylated proteins, estimations reach from 10% to 30% of all proteins phosphorylated at any given time [5, 6], the enrichment of phosphorylated proteins or peptides is of general interest for researchers working on phosphorylation. Indeed, the substoichiometry of phosphorylated protein species (i.e., the low number of phosphorylated proteins compared to their nonphosphorylated counterparts) in complex mixtures is a major analytical challenge in protein phosphorylation analysis [7], and it often provides the basis for research on *in vivo* phosphorylation. Because of its high importance, much work has been dedicated to the development and improvement of methods for phosphopeptide and to a lesser extent for phosphoprotein enrichment in the last years. Novel or refined strategies for phosphopeptide and phosphoprotein enrichment are published continuously. In general, there are two alternatives for the enrichment, namely the enrichment of phosphorylated peptides, or of intact phosphoproteins.

Phosphopeptide enrichment has the advantage that the resulting peptide mixture is less complex compared to a peptide mixture representing a total phosphoprotein digest following phosphoprotein enrichment. However, when enriching for phosphorylated peptides, some drawbacks are recognizable. First, the number of molecules on the peptide level is necessarily much higher than on the protein level. Second, the identification of the respective protein often relies on only one phosphorylated peptide. Finally, studies following the peptide enrichment step are restricted to sequence analysis, and no further experiments on the structure or behavior of the proteins are possible.

In general it is acknowledged that protein identifications by MS/MS should be based on more than one peptide [8]. This is especially true for studies where peptide identifications are achieved using database search, since this is more a homology search than a stringent identification process. The correct identification of phosphorylated peptides via MS is often not as straightforward as is the identification of

unmodified peptides [9]. Therefore, it seems astonishing that, compared to general proteomics studies, less stringent criteria are sometimes applied in phosphoproteomic studies. For robust identification of a phosphoprotein, at least one unmodified peptide of the respective protein in addition to the phosphorylated peptide would be helpful. Alternatively, the protein might be identified by means other than MS (e.g., recognition by an antibody) to support the identification. However, because many researchers try to identify as many phosphoproteins as possible in a given sample, this procedure is not followed in most large-scale phosphoproteomics studies, and identification is often based on a single phosphopeptide per protein.

Enrichment of phosphorylated proteins, in contrast, leads to a compromise between reduction of sample complexity and reliable protein identification when compared to phosphopeptide enrichment or no enrichment. Compared to phosphopeptide enrichment, phosphoprotein enrichment has to deal with different hurdles. First of all, peptides are far more soluble than proteins. Therefore, a much higher number of solvents can be used during the enrichment procedure. The acceptable pH range for peptides spans from near 0 to 14, but β elimination of the phosphate group from Ser and Thr can occur at pH higher than 12. For most proteins the pH at which they are soluble and stable is in the range from pH 6 to 12. In addition, the relative amount of phosphate is usually much higher in a peptide, and the phosphate residues are often better exposed compared to a whole protein. Those characteristics make phosphate affinity purification far easier for peptides than for proteins.

Phosphoprotein Enrichment by Antibodies and Capture Molecules

Antibodies directed against pSer, -pThr [10], and—pTyr [11] have been used to generally enrich for plant proteins phosphorylated at the respective residues in immunoprecipitation experiments. It should be noted that the use of global antibodies directed against pSer or pThr is usually associated with lower specificity compared to the anti-pTyr antibodies. This might be due to the better accessibility of the phosphate group located on Tyr or because of the lower complexity of the structural nature of the Thr and Ser antigens compared to Tyr. Antibodies directed against regions surrounding a specifically phosphorylated residue can be effectively used to enrich for a certain protein of interest. In addition to antibodies, there are several different protein domains designed by nature, which are capable of binding phosphorylated proteins and peptides. While SH2 (Src Homology 2) and PTB domains recognize predominantly Tyr phosphorylation, 14–3–3 proteins as well as WW domains bind to proteins phosphorylated on Ser and Thr, and FHA domains show a preference toward Thr phosphorylation [12]. In principle, all of these domains or the respective proteins can be used for the enrichment of different sorts of phosphorylated peptides. In fact, immobilized 14–3–3 proteins have been used for the purification of Ser/Thr phosphorylated plant proteins [13, 14]. As described, each of these domains/proteins only recognizes a subset of the total pool of phosphorylated proteins, and experimental setups relying on these domains are therefore not as widely applicable as methods relying on more general affinity mechanisms. A combination of a more general affinity chromatography followed by different more restrictive approaches could, however, further facilitate the analysis of the phosphoproteome.

Strategies Relying on Chemical Derivatization

Besides the nonmodifying purification procedures described so far, methods exist that make use of chemical derivatization strategies for the enrichment of phosphorylated proteins and peptides. Two main approaches have been described. The first one consists of β elimination followed by Michael addition, where the phosphate group of pSer or pThr (pTyr cannot be derivatized using this method) is replaced with a nucleophile that is suitable for selective enrichment [15–21]. Alternative strategies involve temporary carbodiimid coupling of the phosphate residues of pSer/pThr/pTyr to a solid phase, followed by washing steps to remove nonphosphorylated species [22, 23]. However, due to the involved chemical processes, unwanted side reactions may occur. This is known for the approach involving β elimination, which is carried out at basic conditions (pH 12–14). These reaction conditions can lead to the replacement of *O*-glycosides [24, 25], sulfonated residues [26], and even of hydroxyl groups, which are located on Ser and Thr [27]. Several other artificial protein modifications are known to occur under highly basic conditions [28]. The carbodiimid approach has the distinct advantage of including pTyr in the analysis and that no side reactions are reported so far. The initial approach [27] involved several modification steps which increases the likelihood of sample loss and modification in each step. The refined method [23, 29] involves methylesterification, which has also its drawbacks as described above (see IMAC section). Nevertheless, the method seems to be promising and might find its way into plant phosphorylation research.

IMAC. The most widespread method used for phosphopeptide enrichment is IMAC. This technique was initially developed by Porath et al. [30] and was originally used to separate all kinds of different proteins. It mainly relies on the attraction of a negatively charged amino acid residue to a positively charged metal that is immobilized on a metal chelator matrix such as iminodiacetic acid. Iron and Ga^{3+} are the most widely used metals for the enrichment of phosphorylated species. Since a phosphate group has a stronger net negative charge than any other amino acid residue, phosphorylated proteins/peptides are better retained on the matrix than their nonphosphorylated counterparts. IMAC has been used to separate plant-derived phosphorylated proteins [31] as well as phosphopeptides [32, 33] from non-phosphorylated ones. However, there are some difficulties when applying IMAC. Histidine and Asp phosphorylations are not accessible by classic IMAC enrichment because they are acid-labile and IMAC loading is usually conducted under acidic conditions. IMAC has in fact been used to enrich for acidic proteins and is prone to unspecific binding of proteins/peptides rich in glutamic and aspartic acid residues [34, 35]. To circumvent the unspecific binding problem, peptides can be methylesterified, thus converting glutamic and aspartic acid residues into their noncharged methyl esters. Unfortunately, this reaction is conducted under harsh conditions (exclusion of water at pH 0–1) and is not free of side reactions such as deamination of glutamines and asparagines [36], making this method of methylesterification especially unsuitable for complex protein mixtures. However, it seems that the success of IMAC and methylesterification also depends on the kind of protein sample and IMAC resin used [37].

Since IMAC has been used for the enrichment of phosphorylated proteins [38] and peptides, it can in principle be applied to couple the advantages of both strategies. However, probably due to experimental difficulties, there is only little literature describing a double enrichment on both the protein and peptide level. A recent description comes from Collins et al. [39], who published a detailed protocol for this sort of sequential purification based on IMAC, and an analogous approach is described by Huang et al. [40]. There is no study describing this or a similar experimental setup for plant proteins. Nevertheless, recent advances in method refinement [see reference 41] might help to overcome some of the aforementioned difficulties of IMAC.

MOAC. Other inorganic approaches make use of metal oxides and hydroxides. This includes the recently developed methods based on zirconia, titania, and aluminum hydroxide ($\text{Al}(\text{OH})_3$). Since oxides and hydroxides are closely related and in fact all of these compounds belong to the oxide class of minerals [see reference 42], the general term metal oxide affinity chromatography may be used for these methods in analogy to IMAC. An example is shown in Figure 29.1.

Titania and Zirconia. Recently, it was shown that metal oxides can also be used for the enrichment of phosphorylated peptides. In these studies, affinity chromatography based on titania (titanium dioxide, TiO_2) [43–52] and zirconia (ZrO_2) [53] was used. On-line coupling of a titania precolumn and an anion exchange [49, 50] or RP column [46, 48] in an HPLC setup has been shown to be useful in the selective analysis of phosphorylated peptides derived from proteolytic digests. Identification of the phosphopeptides was achieved by monitoring the UV trace [49, 50] or by using a mass spectrometer [46, 48]. Similarly, nanoparticles composed of $\text{Fe}_3\text{O}_4/\text{TiO}_2$ core shell particles were used to specifically isolate and detect phosphopeptides [51]. Unspecific binding, which is also reported when using titania [46, 48], can be reduced by methylesterification [48] and the use of appropriate incubation buffers [52]. However, the use of a special incubation buffer is preferable because of the side reactions occurring during methylesterification (see above). While all of the studies named so far focused on the analysis of standard proteins or animal tissue, a very recent application comes from the plant field. In this study the authors used titania to enrich phosphopeptides of spinach stroma membranes and identified some new phosphorylation sites in photosynthesis-related proteins [53].

Aluminum Oxide and Hydroxide. $\text{Al}(\text{OH})_3$ is a well-known and widely used adjuvant in medicine. Adjuvants are additives that enhance the effectiveness of medical treatment by potentiating the immune response and functioning as a carrier for antigens [54, 55]. Because of this wide application, protein adsorption to $\text{Al}(\text{OH})_3$ has been studied in considerable detail and some of those studies also investigate the binding behavior of phosphorylated proteins to $\text{Al}(\text{OH})_3$ [56–59]. In addition, aluminium oxide, a close relative of $\text{Al}(\text{OH})_3$, has been reported to exhibit a high and selective attraction to phosphorylated biomolecules [60]. The conclusion of these studies is that $\text{Al}(\text{OH})_3$ and oxide both have a high affinity for phosphorylated biomolecules and

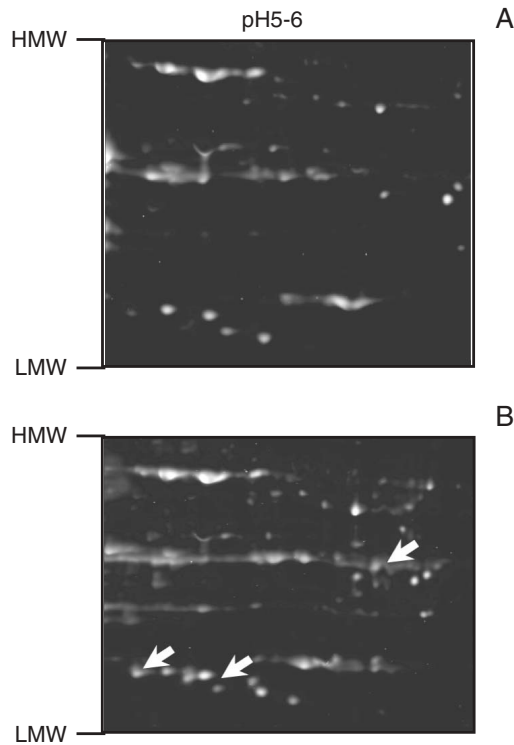


FIGURE 29.1. Phosphoprotein changes of *A. thaliana* after chemical treatment combined with MOAC enrichment: (A) Phosphoproteins of *A. thaliana* cell cultures without treatment. (B) Phosphoproteins of *A. thaliana* after 24 hr 2-deoxyglucose treatment, changes are indicated by white arrows. Proteins are analyzed by LC–MS/MS. Phosphoproteins were visualized with Pro-Q diamond stain. (Chen and Weckwerth, unpublished data.)

that the affinity towards, phosphate is considerably higher than for sulfate or nitrate. These characteristics make them good candidates for the enrichment of phosphorylated proteins and peptides out of complex mixtures. However, it is also clear from some of those studies that nonphosphorylated proteins also exhibit a considerable affinity to those matrices even if they are not phosphorylated.

The enrichment procedure is usually performed under denaturing conditions to decrease the threat of dephosphorylation and degradation events because of the action of endogenous proteases and phosphatases [61, 62]. Proteins are extracted using a mix of buffer, phenol, NaF (which serves as a phosphatase inhibitor), and 2-ME. After precipitation, proteins are resuspended in a special incubation buffer containing urea, subjected to affinity chromatography, and eluted using pyrophosphate. So far, over 300 putative phosphoproteins from different sources including *C. reinhardtii*, *A. thaliana* cell cultures, seeds, and leaves have been enriched (Wolschin and Weckwerth, unpublished results).

29.3 DETECTION OF PROTEIN PHOSPHORYLATION AND DETERMINATION OF PHOSPHORYLATION SITES

In this section, techniques for the general detection of protein phosphorylation and for phosphorylation site pinpointing are described and discussed. All techniques have been applied to plant material. The general detection techniques comprise the following: labeling with radioactive ^{32}P , labeling with a fluorescent probe, detection using anti-phosphoamino acid antibodies, phosphatase treatment in combination with mass shift analysis, and finally MS-based methods alone. Likewise, there is a variety of methods that can be used for protein phosphorylation site pinpointing. Site-specific protein phosphorylation analysis was classically performed using ^{32}P labeling or chemical derivatization in combination with Edman sequencing. The onset of biological MS led to the interpretation of specific fragmentation patterns characteristic for phosphorylated peptides using different instrumental setups.

Antibodies

General phospho-specific and phosphorylation site-specific antibodies can, in concert with an appropriate secondary antibody or directly coupled to an enzyme/fluorescent stain, be used to detect phosphorylation. This method has been successfully applied to detect Tyr, Ser, and Thr phosphorylation in plants [63–66].

Strategies Relying on Chemical Derivatization

Peptides phosphorylated on Ser and Thre can undergo β elimination during fragmentation in a mass spectrometer. This neutral loss of H_3PO_4 often leads to incomplete fragmentation of the peptide in MS-based analyses and complicates the interpretation of spectra. To circumvent this problem, the HPO_3 group can be replaced with a more stable residue, resulting in more informative fragmentation spectra. To this end, β elimination of H_3PO_4 followed by Michael addition can be used for peptides phosphorylated on Ser and Thr. This technique has been used to map phosphorylation sites in plant proteins [67–69].

Radioactive Labeling

For a long time, labeling with radioactive ^{32}P has been the method of choice due to its sensitivity and selectivity. Usually, cell cultures or parts of plants are fed with ^{32}P -orthophosphate *in vivo*. Alternatively, radioactive γ -ATP is applied in *in vitro* labeling experiments using purified enzymes and/or substrates [70]. These procedures are followed by the extraction of proteins and detection of phosphorylation using scintillation counting and/or phosphor-imaging. ^{32}P labeling is used not only to detect phosphorylation but also to identify the peptide and ultimately the amino acid, which is phosphorylated. This can be achieved by Edman sequencing and MS as discussed below. Even though ^{32}P labeling is still the most sensitive approach for the detection of phosphorylation, it has two major drawbacks. First, there are handling and

waste disposal complications since ^{32}P is a radioactive compound and stringent safety rules apply for the use of it. The second and even more concerning issue is that the incorporation of radioactive phosphor may considerably alter the *in vivo* state of the cells compared to nonradioactive phosphor as has been shown in studies involving mammalian cells [71, 72].

Phosphatase Treatment

The treatment of phosphorylated proteins/peptides with a phosphatase results in dephosphorylation accompanied by a mass shift of 80 Da. This approach has been described for different plant phosphoproteins and is a good means to detect and confirm phosphorylation if sample amount is not a limiting factor (one analysis before and one after phosphatase treatment is performed). For example, phosphorylation sites identified in proteins of the moss *P. patens* were confirmed using phosphatase treatment and MS [33]. Detection of phosphorylation of the bifunctional enzyme 6-phosphofructo-2-kinase/FBP 2-phosphatase from *A. thaliana* was supported by alkaline treatment of the whole complex followed by the occurrence of a mass shift after SDS-PAGE [73]. However, phosphatase treatment in combination with mass shift measurements is only a rough proof of phosphorylation.

Edman Sequencing

This technique developed by Pehr Edman led to a major breakthrough in biotechnology. For the first time, the amino acid sequence of a protein could be elucidated. Edman sequencing relies on the sequential degradation of the amino acids at the N-terminus of a polypeptide chain, which is coupled to a solid phase. The reagent phenylisothiocyanate (PTC) is added to the coupled peptide and the modified N-terminal amino acid can be selectively detached under strongly acidic conditions. The resulting amino acid derivatives are analyzed using HPLC. By comparing the elution behavior of the sample and standard amino acid derivatives, stepwise peptide sequencing can be performed. If a phosphopeptide sample was marked with a ^{32}P phosphorylation site, pinpointing can be achieved by concomitant measurement of ^{32}P activity [74]. The major drawbacks are that the proteins to be sequenced have to be purified almost to homogeneity in advance and that it is time-consuming.

Prediction Programs/Phosphorylation Site Databases

There are several phosphorylation site prediction programs available, which rely on different algorithms to elucidate the probability of phosphorylation on specific amino acids for any given protein. However, because of the enormous complexity of the cellular proteome, prediction programs still tend to lead to FP results and their output has to be handled with care [75]. Also, several databases have been constructed, which contain data on experimentally verified phosphorylation sites like ProMEX [76] or PlantsP (<http://plantsp.genomics.purdue.edu/>). As these databases are growing, they are becoming more and more useful for the researcher and they will probably play an important role in the future for serving as encyclopaedias of phosphorylation sites and for the development of more sophisticated prediction programs.

Mutational Analysis

Putative or experimentally identified protein phosphorylation sites can be mutated to confirm phosphorylation and/or to study its biological impact. To this end, mutants are generated with a single amino acid substitution at the phosphorylation site and compared to the wild-type form [77]. This technique is very useful for functional studies of experimentally established phosphorylation sites. However, due to its time-consuming and laborious nature, it is not generally employed in the search for unknown phosphorylation sites.

Dye Technology

Dyes have been recently developed, which selectively stain phosphorylated proteins and peptides no matter if they are phosphorylated on Ser, Thr, or Tyr [78–80]. These dyes rely on the recognition principle of inorganic HPO_3 receptors developed by chemists [81–83]. The structure of the commercially available Pro-Q stain is not public but likely to be similar to the published structures. This dye seems to be very selective when a detailed protocol is stringently followed. It should be noted that *O*-sulfonation can also contribute to the staining, albeit apparently to a lesser extent [78]. A big advantage of these stains is their compatibility with enzymatic in-gel digestion and MS. A protein of interest can be subjected to gel electrophoresis, preliminarily examined for its phosphorylation, digested, and finally analyzed by MS. The disadvantages are the failure to differentiate between the three phosphorylated residues and the lower sensitivity when compared to ^{32}P labeling and antibody-based approaches. However, fluorescent staining was found to greatly facilitate phosphoprotein analysis.

Fragmentation Techniques in MS

Depending on the instrument type used, there is a variety of different fragmentation signatures typical for phosphopeptides, which can be used for the detection of phosphorylation and/or for the determination of phosphorylation sites. Every fragmentation technique leads to different fragmentation patterns, some of them leaving the phosphorylation intact, while others evoke the detachment of the phosphate group during the fragmentation process. Since MS-based methods rely on the determination of m/z , the similarity of the mass shifts caused by phosphate ($+\text{HPO}_3$) and sulfate ($+\text{SO}_3$) ester formation (the difference is only 0.009 mass units) is of concern. Only high-precision instruments like FTICR-MS are able to distinguish between these two compounds following the m/z values.

However, it was reported that CID fragmentation of peptides phosphorylated at Ser or Thr residues leads to a loss of about 98 Da (corresponding to H_3PO_4), while the sulfated counterparts lose 80 Da (corresponding to HSO_3^-) [26]. This special behavior of phosphopeptides can be used to determine if a peptide is phosphorylated on a Ser/Thr or not by searching for a loss of 98 Da (H_3PO_4) in the positive or for 79 Da (PO_3^-) in the negative mode [84–86]. Phosphorylation on Tyr residues is usually more stable, but a loss of 80 Da (HPO_3) in the positive mode or presence of the immonium ion of pTyr (216.043 m/z) is a good indicator of Tyr phosphorylation.

However, one should be aware of the fact that peptides sulfated at Tyr might also lose 80 Da (SO_3^-) in the positive mode [26, 87] and that the nominal mass of 216 Da can also be generated by other ions [88]. CID and post-source decay (PSD) are the most common fragmentation techniques. The remaining techniques, however, promise to be important in future research projects involving plant phosphoproteomics.

PSD. PSD is a fragmentation technique typical for MALDI–TOF–MS instruments. It relies on metastable fragmentation processes proceeding in the flight tube when the instrument is used in the reflector mode. However, this kind of fragmentation often results in spectra that are hard to interpret; and loss of phosphoric acid and related compounds from phosphopeptides is also commonly observed, complicating the identification of phosphorylation sites.

Electron Capture Dissociation (ECD) and Electron Transfer Dissociation (ETD). ECD and ETD are similar fragmentation techniques, which have the big advantage that PTMs such as phosphorylation stay intact during the fragmentation process because of a milder fragmentation procedure [89–92]. ECD was introduced by McLafferty and co-workers [93] and relies on the exposure of multiply protonated peptides to electrons in an FTICR mass spectrometer. The capture of electrons is associated with the fragmentation of the peptide mostly into c- and z-type ions. Unfortunately, ECD is only possible on the very expensive FTICR instruments and not on more affordable mass spectrometers like ion traps, since the molecules have to be held in a special energy stage, which is not easily achieved in ion traps. To overcome this limitation, Syka et al. [91] developed a special interface for ion trap mass spectrometers that facilitates a process which is comparable to ECD. This process is called ETD and also involves the transfer of electrons to protonated peptides. In this technique, radical anionic species of polyaromatic hydrocarbons like fluoranthene are introduced into the trap and transfer electrons to multiply protonated peptides. This transfer leads to an internal fragmentation process believed to be equivalent to the one when using ECD. A slightly modified process can be used to sequence larger peptides of more than 20 amino acids and even proteins [92].

Unfortunately, these techniques are still not widely distributed because suitable instrumentation is very expensive (ECD) or only rarely available (ETD). However, they hold great promise for further research on protein phosphorylation.

CID/MS3 Fragmentation. CID is the most commonly applied fragmentation technique and applies to instrument types such as ion traps and triple quadrupole mass spectrometers. Peptide precursors are protonated on the peptide backbone, recorded (MS), and fragmented using a collision gas like He or Ar (MS/MS or MS2). It is in this fragmentation step that the characteristic loss of 98 for H_3PO_4 can be observed. These fragments can then be fragmented further to yield even more detailed and reliable information of the peptide sequence (MS/MS/MS or MS3). Besides the many advantages the CID-based method has, there are also potential pitfalls that have to be considered. The most important issue, is whether it is true that the dominant loss of 98 Da is indeed always restricted to phosphorylated peptides. To clarify this important issue, the MS2 and MS3 spectra have to be examined in detail.

Quantification of Protein Phosphorylation

While the detection of phosphorylation and the identification of phosphorylation sites is highly important to get a first glance at the phosphorylation state of proteins and delivers the fundament of phosphorylation research, quantification of these phosphorylation events gives an even deeper insight into complex protein regulation. This quantification can be performed on the proteome, protein, peptide, and site-specific level. However, the quantification of protein phosphorylation imposes great challenges to the experimenter because of the low abundance of phosphoproteins and the complex fragmentation of phosphopeptides observed in MS/MS-based analyses.

Methods based on ^{32}P labeling or global anti-phospho amino acid antibodies only detect and quantify the phosphorylated version of a protein but leave the unphosphorylated forms undetected, while other methods can also determine phosphorylation stoichiometries (the percentage of phosphorylated proteins). Several approaches have been developed to cope with the inherent difficulties associated with quantification, and these approaches are used to unravel complex protein regulation achieved by phosphorylation/dephosphorylation events.

Imaging Methods. One example is a general approach relying on the attachment of a fluorescent dye to phosphorylated residues followed by an intensity-based measurement of the phosphoprotein/phosphopeptide dye complex. An increase in the number of phosphorylated residues or the phosphorylation stoichiometry results in a measurable increase of the fluorescence [94]. Nonphosphorylated species of the same proteins can be observed and quantified using dyes like SYPRO Ruby or silver stain. An equivalent method is the one that makes use of antibodies. Phosphorylation is visualized using a secondary antibody coupled to an enzyme or a fluorescent dye, and increase or decrease of phosphorylation is determined by the changes in the product produced by the enzyme or by the changes in the fluorescence. Global phospho-specific antibodies as well as antibodies directed against specific phosphorylated antigens and their nonphosphorylated counterparts can be used as primary antibodies in this approach to obtain results on phosphorylation levels. However, the drawbacks of the use of these antibodies, which were discussed in the previous sections, have to be taken into account.

Similarly, a change in phosphorylation can be monitored by comparing scintillation counts of two samples labeled with ^{32}P -orthophosphate *in vivo* or with radioactive γ -ATP *in vitro*. Nevertheless, uptake and incorporation of ^{32}P is probably never complete, and endogenous as well as previously bound nonradioactive ^{33}P always presents a problem for quantification. Relative quantification of the global phosphorylation state of a protein or peptide can be obtained readily using these methods and always bearing in mind the limitations of these techniques discussed in previous sections.

However, all of these approaches except the one relying on specific antibodies have the major disadvantage that site-specific information is lacking. This issue becomes especially important when one is looking at a protein or peptide, which can be phosphorylated at multiple residues, and each phosphorylation event may represent another activity state. This is also true for approaches where inductively

coupled plasma (ICP) is used to determine the content of phosphorus in a protein [95, 96]. Therefore, the described methods are only suitable if one is interested in global changes of phosphorylation, for example, in whole proteomes or if one can rule out the possibility of multisite phosphorylation.

29.4 MS-BASED APPROACHES

More recent methods rely on the use of MS because of its selectivity, sensitivity, and capability of direct determination of phosphorylation sites. There is a wide variety of different methods based on MS ranging from methods depending on chemical derivatization to the use of external and internal standards for the investigation of changes in phosphorylation, some of which have been applied to plant protein phosphorylation. While many methods have been developed for the quantification of peptides and proteins in general, descriptions in the following paragraphs are restricted to the ones that were primarily designed for phosphopeptides/phosphoproteins. For a more general discussion on protein quantification, the reader is referred to Ong and Mann [97] and Glinksi and Weckwerth [98].

Strategies Relying on Chemical Derivatization

Overall there are two different kinds of derivatization techniques, which are commonly used for the relative quantification of phosphorylation. The first is based on β elimination and Michael addition. In this approach, two samples (proteins or peptides) to be compared are derivatized with differently deuterated forms (i.e., d0/d5) or different isotopes (i.e., C12/C13) of the same nucleophile, the samples are mixed, proteins are digested, and peptides are subjected to chromatography, which is directly coupled to a mass spectrometer. The ratio of the corresponding peptide pairs leads to a relative quantification of phosphorylation. In the second derivatization strategy, peptides derived from two different samples (control and treated) are derivatized with the undeuterated and the deuterated form of methanol, thus converting carboxyl residues into their corresponding methyl esters in the process of methylesterification. Phosphopeptides are then usually enriched to overcome the low level of phosphorylation and peptide pairs are analyzed by MS as described above. The major drawbacks of these approaches are the side reactions observed in chemical derivatization and elution differences of the peptide pairs in chromatography, depending on the type of derivatization agent. Especially deuterated forms of the peptides are known to exhibit a slight retention time shift in chromatography when compared to their nondeuterated counterparts, and this sometimes impairs direct comparison of peptide pairs [99]. Also, the number of samples that can be compared is limited due to the limited availability of differentially deuterated or isotopically coded tags.

Other strategies involving stable isotope labeling by metabolic labeling or tagging (like SILAC or iTRAQ) are not primarily applied to phosphoproteins/peptides and are not further discussed [for reviews see references 97 and 100].

Methods Based on ICP-MS

ICP-MS is a technique that is capable of analyzing atoms instead of molecules. A very powerful implementation of this technique was presented by Lehmann and co-workers [101–103], who used peptide separation via HPLC directly coupled to ICP-MS to assess the global phosphorylation degree in a sample by monitoring the phosphorus and the sulfur ratio. In this approach, peptides are eluted sequentially into the ICP-MS and decomposed into their atoms in the plasma. In real time, sulfur as well as phosphorus traces are recorded and related to each other. We used this approach to monitor differences in the phosphorylation degree during development of *A. thaliana*. Profound differences were observed that were typical for each developmental stage (Figure 29.2).

Methods Involving Standard Peptides

The methods involving internal standards can be divided in two sections: one involving the addition of a standard, which is not derived from the sample [104, 105] and one where peptides derived from the protein of interest are used [106–110]. Integrated ion traces of the phosphopeptide are then standardized using the ion traces of other, nonphosphorylated peptides. These ratios can be compared for multiple samples, and thereby differences in the extent of phosphorylation can be determined.

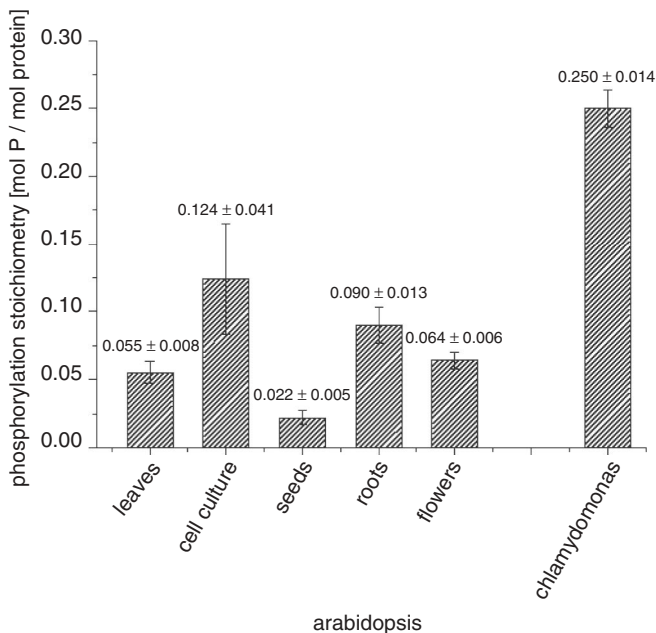


FIGURE 29.2. ICP-MS analysis of total phosphoprotein content of *A. thaliana* and *C. reinhardtii* reveals tissue-dependent composition of the phosphoproteome.

Since these methods are noninvasive in that they do not require chemical modification of the protein of interest, they are especially well-suited to disclose the native phosphorylation of a protein.

29.5 BIOLOGICAL IMPLICATIONS OF PROTEIN MULTISITE PHOSPHORYLATION IN PLANTS

Some of the methods discussed so far have been used to gain insight into the variable and complex nature of protein multisite phosphorylation in plants. While it is not the intention of the authors to provide a complete overview of multisite protein phosphorylations detected in the plant field so far, we want to provide examples that give insight into the variety of cellular mechanisms that can be influenced by this important regulatory pathway.

Signaling/PPI

Phosphorylation has been observed in a wide variety of plant proteins and is described to play a role in different aspects of protein regulation. An important process that depends on phosphorylation in plants is phytochrome signaling. Phytochromes are proteins mediating biochemical processes in response to light. They are produced in an inactivated form and are activated upon light adsorption followed by a conformational change. Signals are then transduced either indirectly or directly by binding of the activated phytochrome to transcriptional activators. Three phosphorylation sites have been identified on phytochrome using different methods including radioactive labeling, Edman degradation, mutation analysis, and MS [67, 111–113]. Phosphorylation of the activated form is described to attenuate light signaling processes, while dephosphorylation leads to enhanced photoresponsiveness [114]. Phosphorylation is reported to inhibit PPIs, between phytochrome and its putative transducer proteins NDPK2 and PIF3 (a basic transcription factor) [115], thereby interrupting the signal transduction process.

Another protein class widely involved in signal transduction processes is formed by the CDPKs. These proteins are composed of an N-terminal domain, a catalytic domain, a junction (j) domain that can autoinhibit the kinase, and a CaM-like domain responsible for calcium binding. The binding of calcium to the CaM-like domain exerts the release of the autoinhibitory function of the j domain. Recently, a comprehensive study of CDPK phosphorylation was published by Hegeman et al. [116]. In this study, several CDPKs were overexpressed *in vitro* and their capability of autophosphorylation was examined by MS. In total, 31 new phosphorylation sites were mapped on 7 CDPKs and 2 CDPK-related kinases (CRKs), thus suggesting the importance of CDPK multisite phosphorylation. Other signal transduction processes are also directly linked to phosphorylation. BRs, which regulate plant growth and development, are reported to lead to multiple phosphorylations of BRI1 and BAK1 as was suggested by IMAC coupled to LC-MS/MS and partly supported by mutation analysis [117]. Phosphorylation of these proteins seems to be important for their kinase activity and might be linked to heterodimerization.

Symbiosis

A special case of signal transduction is the one involved in interspecies contact. The establishment of symbiosis with nitrogen-fixing rhizobia and phosphate acquiring AM fungi was shown to depend on multisite phosphorylation by Yoshida and Parniske [118]. By mutation analysis and MS, they determined phosphorylation sites in SYMRK, which is a RLK crucial for symbiosis in plants [118, 119]. Activation of this kinase depends on the phosphorylation of at least two phosphorylation sites, and good indications exist for the involvement of a third site.

Photosynthesis

Phosphorylation also plays a major role in photosynthetic and associated processes. The thylakoid-soluble protein of 9 kDa (TSP9) was characterized as a phosphoprotein by Carlberg et al. [120]. They identified three phosphorylated Thr(s) using MS and showed that redox-dependent phosphorylation of these sites leads to release of the protein from the membrane. Many other proteins involved in photosynthesis were also shown to be phosphorylated; but for most of them, only single phosphorylation sites are known [107, 108]. However, Vener et al. [107, 121] described a light-dependent phosphorylation of PsbH (PS II reaction center H protein) at two Thr(s), and phosphorylation of PsbH was reported to increase its stability². In a large-scale study of *C. reinhardtii* thylakoid membranes conducted by Turkina and co-workers, several photosynthesis-related proteins were shown to be phosphorylated following an experimental setup consisting of membrane purification, peptide methylation, phosphopeptide enrichment by IMAC, and MS-based analysis [122]. Interestingly, it could be shown that the extent of phosphorylation varied for several peptides, depending on physiological conditions.

Cell-to-Cell Trafficking

Selective phosphorylation of the so-called TMV movement protein (TMV-MP) also influences the spreading of tobacco mosaic virus in some tobacco species. Phosphorylation sites were identified by replacing putative phosphorylation sites in recombinant movement protein with alanine and glycine and subsequent determination of the absence or presence of phosphorylation using radioactive ATP *in vitro* [123]. Trutnyeva and co-workers [124] then replaced the identified Ser and Thr phosphorylation sites in a sequential manner with negatively charged amino acids and compared virus spreading for each mutant. While the phosphorylation of TMV-MP at one specific Ser seems to stimulate cell-to-cell movement of the virus, phosphorylation of two other sites has the opposite effect. The authors conclude that this multisite phosphorylation event may proceed in a sequential manner: First, phosphorylation leads to a spreading of the virus (a benefit for the virus), but further phosphorylation events then restrict the harmful infection (a benefit for the host plant).

Growth Control

Phosphorylation of ribosomal protein S6 at several amino acids at its C-terminus end is well known to be dependent on insulin/IGF signaling in animals [125]. It is required for the activity of S6 leading to the translation of 5' TOP mRNAs (mRNAs that contain an oligopyrimidine tract at their 5' transcriptional start site) coding for proteins of the protein synthetic apparatus. In plants, multisite phosphorylation of specific sites in S6 in analogy to the animal system has been reported for maize using ^{32}P labeling and MALDI-TOF-MS [126]. Here, following phosphoprotein enrichment and MS-based peptide analysis, we could identify an S6 phosphopeptide in *A. thaliana* belonging to the C-terminal end of the protein. Phosphorylation of *A. thaliana* S6 ribosomal protein was also reported by Chang et al. [69] who subjected enriched ribosomes to β elimination and MS. Furthermore, studies exist describing aspects of a signal transduction pathway in plants similar to the animal pathway leading from membrane perception of mitogens to S6 phosphorylation and growth control [126–131]. Thus it seems that this important pathway, including multisite phosphorylation on specific residues at the C-terminus of S6, is indeed conserved in plants.

Enzymatic Activity

Of high importance and well-described is the impact of protein phosphorylation on enzymatic activity. A classic example is PDH, which forms part of the PDH complex (PDC). This key enzyme, which converts pyruvate to acetyl-CoA and produces NADH, has been shown by mutagenesis approaches and radioactive labeling to be phosphorylated in mammals on multiple serines. Phosphorylation leads to deactivation of the enzyme and might play a role in the formation of the PDC [for review see reference 124]. In plants, however, the situation is more complex since they possess not only a mitochondrial PDC but also a plastidic PDC. No phosphorylation sites are reported for plastidic PDC so far, but two sites (one homologous to mammalian site 1 and a second Ser one upstream of the mammalian site 2) have been confirmed in pea by MS [132]. Phosphorylation of plant PDH leads to its inactivation as in the case of the mammalian PDH [133]. Future studies are likely to reveal the exact regulation of plant mt PDH by multisite phosphorylation.

Bykova et al. [134] found multisite phosphorylation of formate dehydrogenase in potato tubers using a combination of ^{32}P *in vitro* labeling, IMAC, and MS. They showed that phosphorylation led to deactivation of the enzyme and phosphorylation was inhibited by the addition of NAD⁺, formate, and pyruvate in a manner similar to that of PDH. However, no detailed studies about the functional influence of individual phosphorylation were performed.

One of the most extensively studied examples for plant protein phosphorylation is sucrose-phosphate synthase (SPS). This metabolic key enzyme catalyzes the formation of sucrose-phosphate from fructose-6-phosphate and UDP-glucose. Three phosphorylation sites have been described for SPS in spinach and tobacco using a wide variety of methods including radioactive labeling, site-directed mutagenesis, and MS. Phosphorylation clearly influences the activity of this protein. One phosphorylation site is located in a 14–3–3 binding domain; another is responsible for light/dark modulation,

and a third site has been reported to be involved in osmotic stress response [135–138]. Interestingly, phosphorylation of the light/dark modulation site is responsible for deactivation of the enzyme, while phosphorylation of the osmotic stress response site leads to the opposite effect. Very recently, using a MS approach based on internal standard peptides, we could show that the phosphorylation dynamics of this important enzyme in *A. thaliana* are likely to be temperature dependent [110].

Another important enzyme of plant sugar metabolism, trehalose-6-phosphate-synthase (TPS), was also recently shown to be multiply phosphorylated. TPS catalyzes the formation of trehalose-6-phosphate (T6P), which is suggested to be a sugar signal that has a profound effect on nutrient assimilation and development. Glinski and Weckwerth [98] used an MS-based approach to confirm putative phosphorylation sites in several different TPS isoforms. The sites were predicted on the basis of a recognition motif described by Hardie et al. [139], and peptides including these sites were subjected to *in vitro* phosphorylation followed by MS analysis. The probability of truly important phosphorylation events was estimated by the level of phosphorylation the peptides exhibited. Also, Harthill and co-workers showed phosphorylation-dependent binding of TPS to 14–3–3 proteins and after *in vitro* phosphorylation determined two phosphorylation sites by MS [140].

29.6 LARGE-SCALE PROTEOMICS STUDIES

While the examples discussed so far are mostly based on targeted approaches, there are also nontargeted or “omics” approaches that have delivered insight into plant protein multisite phosphorylation. These studies aim at the identification of many phosphorylation sites in a single analysis and often lead to the finding of novel multisite protein phosphorylations. Recently, we developed a strategy for phosphoprotein enrichment using $\text{Al}(\text{OH})_3$ and neutral loss scanning of H_3PO_4 for phosphopeptides identification out of complex samples [138]. Other studies use subcellular fractionation techniques [32] or treating tissue *in vivo* with phosphatase inhibitors long before protein extraction to elevate the level of phosphorylation [141]. The criteria for phosphopeptide identification vary between studies making a direct comparison difficult. The choice of useful filter criteria is an ongoing debate of researchers [8, 142, 143]. Manual validation and the application of stringent filtering criteria are often the only way to minimize false-positive rates. However, the data obtained in more comprehensive investigations of protein phosphorylation in living organisms are crucial and can provide the basis for in-depth studies of the identified phosphorylation sites. Enrichment of phosphorylated proteins and peptides is a common theme in this kind of studies because it overcomes the low abundance of these biomolecules. By building up phosphorylation site libraries, the research community can considerably increase the number of putatively interesting targets. These libraries can be used to study their *in vitro* phosphorylation in detail using methods such as a recent approach exploiting peptide-multiple reaction monitoring on a triple quadrupole mass spectrometer [144].

29.7 CONCLUSIONS AND FIVE-YEAR VIEWPOINT

Broad-scale research on protein multisite phosphorylation in plants is still in its infancy. However, examples exist describing various aspects of the importance of protein multisite phosphorylation as a regulatory mechanism. In nature, in addition to protein phosphorylation other protein modifications contribute to the complex signaling pathways. Thus, phosphorylation is only one, but very important, layer of PTM influencing interaction and regulation of proteins.

While it is still not as sensitive and selective as ^{32}P labeling, fluorescent dye technology has the potential to replace radioactive methods in the future. Because it is not invasive and is applied after protein extraction from their *in vivo* sources, it does not alter the phosphorylation state of proteins in a consequence of the experimental setup as might be the case when using ^{32}P . MS-based methods are increasingly becoming indispensable for the study of protein phosphorylation because of their sensitivity and selectivity. The neutral loss-driven MS3 approach, which is increasingly becoming popular, is a good means to obtain high-quality data on protein phosphorylation, qualitatively as well as quantitatively, when considering the MS2 data. In the future, even higher sophisticated instruments will be available and MS-based research will probably become part of any laboratory working on protein biochemistry. Mutation/knockout analyses will continue to be important when specific functions are to be assigned to site-specific protein phosphorylations.

Further research on protein multisite phosphorylation is likely to detect important regulatory mechanisms and signaling pathways in plants; and by comparison to the better-studied animal models, it can reveal major similarities and differences between the different kingdoms of life. Most work is restricted to phosphorylation events on Ser and Thr because these are regarded as the most important phosphorylation events in plants. However, it should be noted that His/Asp phosphorylation is also reported for plants and is in fact an important event that triggers signal transduction in response to extracellular stimuli including cytokinin and ethylene signaling [145–147]. Therefore, it seems likely that this kind of phosphorylation will be an important research area in the future.

REFERENCES

1. Roach PJ. 1991. *J Biol Chem* **266**:14139–14142.
2. Cohen P. 2000. *Trends Biochem Sci* **25**:596–601.
3. Hunter T and Cooper JA. 1985. *Annu Rev Biochem* **54**:897–930.
4. Sickmann A, Mreyen M and Meyer HE. 2002. *IUBMB Life* **54**:51–57.
5. Alberts B, et al. 1994. In: *Molecular Biology of the Cell*, 3rd ed., Garland Publishing, New York, pp. 195–222.
6. Mann M, et al. 2002. *Trends Biotechnol* **20**:261–268.
7. Steen H, et al. 2006. *Mol Cell Proteomics* **5**:172–181.
8. Taylor GK and Goodlett DR. 2005. *Rapid Commun Mass Spectrom* **19**:3420.
9. Salih E. 2005. *Mass Spectrom Rev* **24**:828–846.

10. Laugesen S, Bergoin A, and Rossignol M. 2004. *Plant Physiol Biochem* **42**:929–936.
11. Novikova GV, et al. 2000. *FEBS Lett* **474**:29–32.
12. Roque AC and Lowe CR. 2005. *Biotechnol Bioeng* **91**:546–555.
13. Milne FC, et al. 2002. *Biochem Soc Trans* **30**:379–381.
14. Kulma A, et al. 2004. *Plant J* **37**:654–667.
15. Goshe MB, et al. 2001. *Anal Chem* **73**:2578–2586.
16. Oda Y, Nagasu T, and Chait BT. 2001. *Nat Biotechnol* **19**:379–382.
17. Goshe MB, et al. 2002. *Anal Chem* **74**:607–616.
18. Qian WJ, et al. 2003. *Anal Chem* **75**:5441–5450.
19. Thaler F, et al. 2003. *Anal Bioanal Chem* **376**:366–373.
20. van der Veken P, et al. 2005. *Chembiochem* **6**:2271–2280.
21. Vosseller K, et al. 2005. *Proteomics* **5**:388–398.
22. Zhou H, Watts JD, and Aebersold R. 2001. *Nat Biotechnol* **19**:375–378.
23. Tao WA, et al. 2005. *Nat Methods* **2**:591–598.
24. Wells L, et al. *Proteomics* **1**:791–804.
25. Amoresano A, et al. 2004. *Eur J Mass Spectrom* **10**:401–412.
26. Medzihradszky KF, et al. 2004. *Mol Cell Proteomics* **3**:429–440.
27. McLachlin DT and Chait BT. 2003. *Anal Chem* **75**:6826–6836.
28. Whitaker JR and Feeney RE. 1983. *Crit Rev Food Sci Nutr* **19**:173–212.
29. Bodenmiller B, et al. 2007. *Mol Biosyst* **3**:275–286.
30. Porath J, et al. 1975. *Nature* **258**:598–599.
31. Anguenot R, Yelle S, and Nguyen-Quoc B. 1999. *Arch Biochem Biophys* **365**:163–169.
32. Nuhse TS, et al. 2003. *Mol Cell Proteomics* **2**:1234–1243.
33. Heintz D, et al. 2004. *Electrophoresis* **25**:1149–1159.
34. Zachariou M and Hearn MTW. 1995. *J Protein Chem* **14**:419–430.
35. Ueda EK, Gout PW and Morganti L. 2003. *J Chromatogr A* **988**:1–23.
36. Rohde P, Hinch DK, and Heyer AG. 2004. *Plant J* **38**:790–799.
37. Haydon CE, et al. 2003. *Mol Cell Proteomics* **2**:1055–1067.
38. Andersson L and Porath J. 1986. *Anal Biochem* **154**:250–254.
39. Collins MO, et al. 2005. *Sci STKE*, pp. 16.
40. Huang SY, Hsu JL, and Chen SH. 2005 *J Chinese Chem Soc* **52**:765–772.
41. Kinoshita E, et al. 2005. *J Sep Sci* **28**:155–162.
42. Strunz H and Nickel E. 2001. E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart.
43. Marsuda H, Nakamura H, and Nakajima T. 1990. *Anal Sci* **6**:911–912.
44. Murayama K, et al. 1994. *Microchem. J* **49**:362–367.
45. Ikeguchi Y and Nakamura H. 1997. *Anal Sci* **13**:479–483.
46. Kuroda I, et al. 2004. *Anal Sci* **20**:1313–1319.
47. Miyazaki S, et al. 2004. *J Chromatography A* **1043**:19–25.
48. Pinkse MWH, et al. 2004. *Anal Chem* **76**:3935–3943.
49. Sano A and Nakamura H. 2004. *Anal Sci* **20**:565–566.
50. Sano A and Nakamura H. 2004. *Anal Sci* **20**:861–864.
51. Chen CT and Chen YC. 2005. *Anal Chem* **77**:5912–5919.

52. Larsen MR, et al. 2005. *Mol Cell Proteomics* **4**:873–886.
53. Kweon HK and Hakansson K. 2006. *Anal Chem* **78**:1743–1749.
54. Rinalducci S, et al. 2006. *J Proteome Res* **5**:973–982.
55. Baylor NW, Egan W and Richman P. 2002. *Vaccine* **20** (Suppl 3):S18–S23.
56. Francis JN and Durham SR. 2004. *Curr Opin Allergy Clin Immunol* **4**:543–548.
57. Kato H, et al. 1994. *Microbiol Immunol* **38**:543–548.
58. Chang MF, et al. 1997. *PDA J Pharm Sci Technol* **51**:25–29.
59. Shi SD, et al. 2001. *Anal Chem* **73**:19–22.
60. Iyer S, HogenEsch H, and Hem SL. 2003. *Pharm Dev Technol* **8**:81–86.
61. Colettipreviero MA and Previero A. 1989. *Analytical Biochemistry* **180**:1–10.
62. Wolschin F and Weckwerth W. 2005. *Plant Methods* **1**:9.
63. Wolschin F, Wienkoop S, and Weckwerth W. 2005. *Proteomics* **5**:4389–4397.
64. Pursiheimo S, et al. 1998. *FEBS Lett* **423**:178–182.
65. Tullberg A, Hakansson G, and Race HL. 1998. *Biochem Biophys Res Commun* **250**: 617–622.
66. Barizza E, et al. 1999. *FEBS Lett* **447**:191–194.
67. Azad AK, et al. 2004. *Plant Cell Physiol* **45**:608–617.
68. Lapko VN, et al. 1997. *Biochemistry* **36**:10595–10599.
69. Gliniski M, et al. 2003. *Rapid Commun Mass Spectrom* **17**:1579–1584.
70. Chang IF, et al. 2005. *Plant Physiol* **137**:848–862.
71. Duff S and Chollet R. 1995. *Plant Physiol* **107**:775–782.
72. Bond JA, et al. 1999. *Oncogene* **18**:3788–3792.
73. Craig AL, et al. 1999. *Oncogene* **18**:6305–6312.
74. Furumoto T, et al. 2001. *Plant Cell Physiol* **42**:1044–1048.
75. Pierce MW, et al. 1982. *J Biol Chem* **257**:10681–10686.
76. Blom N, et al. 2004. *Proteomics* **4**:1633–1649.
77. Hummel J, et al. 2007. *BMC Bioinform* **8**:216.
78. Toroser D, et al. 1999. *Plant J* **17**:407–413.
79. Martin K, et al. 2003. *Proteomics* **3**:1244–1255.
80. Martin K, et al. 2003. *Comb Chem High Throughput Screening* **6**:331–339.
81. Kinoshita E, et al. 2006. *Mol Cell Proteomics* **5**:749–757.
82. Han MS and Kim DH. 2002. *Angew Chem Int Ed Engl* **41**:3809–3811.
83. Han MS and Kim DH. 2003. *Bioorg Med Chem Lett* **13**:1079–1082.
84. Tobey SL, Jones BD, and Anslyn EV. 2003. *J Am Chem Soc* **125**:4026–4027.
85. DeGnore JP and Qin J. 1998. *J Am Soc Mass Spectrom* **9**:1175–1188.
86. Tholey A, Reed J, and Lehmann WD. 1999. *J Mass Spectrom* **34**:117–123.
87. Edelson-Averbukh M, Pipkorn R, and Lehmann WD. 2006. *Anal Chem* **78**:1249–1256.
88. Salek M, Costagliola S, and Lehmann WD. 2004. *Anal Chem* **76**:5136–5142.
89. Steen H, et al. 2001. *Anal Chem* **73**:1440–1448.
90. Stensballe A, et al. 2000. *Rapid Commun Mass Spectrom* **14**:1793–1800.
91. Shi Y, HogenEsch H and Hem SL. 2001. *Vaccine* **20**:80–85.
92. Syka JE, et al. 2004. *Proc Natl Acad Sci USA* **101**:9528–9533.

93. Coon JJ, et al. 2005. *Proc Natl Acad Sci USA* **102**:9463–9468.
94. Zubarev RA, Kelleher NL, and McLafferty FW. 1998. *J Am Chem Soc* **120**:3265–3266.
95. Steinberg TH, et al. 2003. *Proteomics* **3**:1128–1144.
96. Wind M, et al. 2003. *Electrophoresis* **24**:1276–1280.
97. Wind M, et al. 2003. *Anal Biochem* **317**:26–33.
98. Ong SE and Mann M. 2005. *Nat Chem Biol* **1**:252–262.
99. Glinski M and Weckwerth W. 2006. *Mass Spectrom Rev* **25**:173–214.
100. Zhang R and Regnier FE. 2002. *J Proteome Res* **1**:139–147.
101. Guerrero IC and Kleiner O. 2005. *Biosci Rep* **25**:71–93.
102. Wind M, et al. 2001. *Anal Chem* **73**:29–35.
103. Wind M, Wesch H, and Lehmann WD. 2001. *Anal Chem* **73**:3006–3010.
104. Kruger R, et al. 2006. *Anal Chem* **78**:1987–1994.
105. Ruse CI, et al. 2002. *Anal Chem* **74**:1658–1664.
106. Wienkoop S and Weckwerth W. 2006. *J Exp Bot* **57**:1529–1535.
107. Tsay YG, et al. 2000. *Anal Biochem* **287**:55–64.
108. Vener AV, et al. 2001. *J Biol Chem* **276**:6959–6966.
109. Aro EM, Rokka A and Vener AV. 2004. *Methods Mol Biol* **274**:271–285.
110. Steen H, et al. 2005. *Proc Natl Acad Sci USA* **102**:3948–3953.
111. Wolschin F, et al. 2005. *Rapid Commun Mass Spectrom* **19**:3626–3632.
112. McMichael RW, Jr and Lagarias JC. 1990. *Biochemistry* **29**:3872–3878.
113. Lapko VN, Wells TA, Song PS. 1996. *Biochemistry* **35**:6585–6594.
114. Lapko VN, et al. 1999. *Protein Sci* **8**:1032–1044.
115. Rubio V and Deng XW. 2005. *Cell* **120**:290–292.
116. Kim JJ, et al. 2004. *Plant Cell* **16**:2629–2640.
117. Hegeman AD, et al. 2006. *Proteomics* **6**:3649–3664.
118. Wang X, et al. 2005. *Plant Cell* **17**:1685–1703.
119. Yoshida S and Parniske M. 2005. *J Biol Chem* **280**:9203–9209.
120. Stracke S, et al. 2002. *Nature* **417**:959–962.
121. Carlberg I, et al. 2003. *Proc Natl Acad Sci USA* **100**:757–762.
122. Turkina MV, et al. 2006. *Mol Cell Proteomics* **5**:1412–1425.
123. Vener AV, Ohad I, and Andersson B. 1998. *Curr Opin Plant Biol* **1**:217–223.
124. Citovsky V, et al. 1993. *Genes Dev* **7**:904–910.
125. Kolobova E, et al. 2001. *Biochem J* **358**:69–77.
126. Avruch J, et al. 2005. *Curr Opin Clin Nutr Metab Care* **8**:67–72.
127. Williams AJ, et al. 2003. *Plant Physiol* **132**:2086–2097.
128. Sanchez-de-Jimenez E, Aguilar R, and Dinkova T. 1997. *Biochimie* **79**:187–194.
129. Garcia Flores C, et al. 2001. *Biochem J* **358**:95–100.
130. Beltran-Pena E, et al. 2002. *Physiol Plant* **115**:291–297.
131. Reyes de la Cruz H, Aguilar R, and Sanchez de Jimenez E. 2004. *Biochemistry* **43**:533–539.
132. Mahfouz MM, et al. 2006. *Plant Cell* **18**:477–490.
133. Tovar-Mendez A, Miernyk JA, and Randall DD. 2003. *Eur J Biochem* **270**:1043–1049.

134. Rao KP and Randall DD. 1980. *Arch Biochem Biophys* **200**:461–466.
135. Bykova NV, et al. 2003. *J Biol Chem* **278**:26021–26030.
136. McMichael RW, Jr, et al. 1993. *Arch Biochem Biophys* **307**:248–252.
137. Huber SC and Huber JL. 1996. *Annu Rev Plant Physiol Plant Mol Biol* **47**:431–444.
138. Toroser D, Athwal GS, and Huber SC. 1998. *FEBS Lett* **435**:110–114.
139. Winter H and Huber SC. 2000. *Crit Rev Biochem Mol Biol* **35**:253–289.
140. Harthiu JE, et al. 2006. *Plant J* **47**:211–223.
141. Hardie DG, Carling D, and Carlson M. 1998. *Annu Rev Biochem* **67**:821–855.
142. Wagner V, et al. 2006. *Eukaryot Cell* **5**:457–468.
143. Carr S, et al. 2004. *Mol Cell Proteomics* **3**:531–533.
144. Bradshaw RA. 2005. *Mol Cell Proteomics* **4**:1223–1225.
145. Gliniski M and Weckwerth W. 2005. *Mol Cell Proteomics* **4**:1614–1625.
146. Sakakibara H, Taniguchi M, and Sugiyama T. 2000. *Plant Mol Biol* **42**:273–278.
147. Sakakibara H. 2003. *J Plant Res* **116**:253–257.

PROTEOME ANALYSIS OF THE UBIQUITIN PATHWAY

Junmin Peng

30.1 INTRODUCTION

In 1975, a small protein of 76 amino acid residues was found to be ubiquitously present in cells and was therefore named ubiquitin (Ub) [1]. The function of this protein remained a mystery until 1980, when Ub was shown to modify protein substrates covalently and target them for degradation in reconstructed reticulocyte by the group of Hershko, Ciechanover, and Rose [2, 3], who won the 2004 Nobel Prize in chemistry for this fundamental discovery. The concept was further substantiated by the lab of Varshavsky to demonstrate the physiological function of Ub during protein turnover in cells [4, 5]. Since then, Ub has risen as a central regulator in almost all cellular processes [6].

Protein modification by Ub is now known to function as a versatile molecular signal in a variety of cellular events, such as proteasome-mediated degradation, protein sorting, apoptosis, DNA repair, and regulation of transcription and translation [6, 7]. Proteins are usually attached by Ub in three steps in an ATP-dependent manner. Ub is first activated by forming a thioester intermediate with a Ub-activating enzyme (E1), then transferred to a Ub-conjugating enzyme (E2) through a thioester bond, and

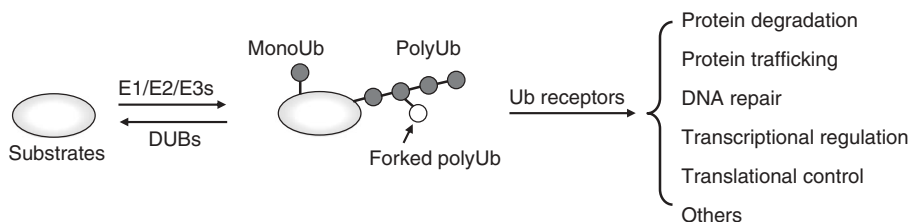


FIGURE 30.1. Major regulatory stages in the Ub signaling pathway. Protein substrates are modified by ubiquitin through coordinated action of E1, E2s, and E3s to assemble multiple forms of Ub chains (monoUb, polyUb, and even forked polyUb). The modifications can be removed or edited by the activity of DUBs. The signal associated with ubiquitination is decoded by UBPs as receptors and propagated in a wide range of cellular processes.

finally attached to targeted proteins by a Ub ligase (E3). In the ubiquitinated proteins, the carboxyl group of the Gly residue at the C-terminus of UB forms an isopeptide bond with the epsilon amino group of a Lys residue of the substrates. In some cases, the N-terminal amino group [8] or Cys residue [9] may also serve as alternative ubiquitination sites in proteins. Moreover, Ub can be conjugated to the targeted proteins in the form of monomers (monoubiquitination) or polymers (polyubiquitination), and the polymers may even grow forked chains (Figure 30.1). As opposed to Ub conjugation, Ub is cleaved from the substrates or free polyUb chains by various deubiquitinating enzymes (DUBs) [10]. In the human proteome, only a few E1 enzymes are characterized and about 50 different E2s are estimated to exist. More than 500 human E3 ligases are proposed to recognize thousands of substrates with specificity [11]. In addition, the specificity in Ub signaling is believed to rely on the interaction between Ub moieties (monoUb or polyUb) with Ub-binding proteins (UBPs) as receptors [12] (Figure 30.1). To date, about 15 families of Ub-binding domains (UBDs) in many different sequence contexts (>250 human proteins) were identified to interact with Ub. Other than Ub, a number of Ub-like proteins (ULPs) have been discovered to display similar 3D structure and to covalently modify target proteins and even lipids, such as SUMO (small Ub-related modifier), NEDD8 (neural precursor cell-expressed developmentally down-regulated gene 8), ISG15 (interferon-stimulated gene 15), and Atg12 and Atg8 (*autophagy-related gene 8*) in distinct functional pathways [13].

30.2 BRIEF BIBLIOGRAPHIC REVIEW

The development of modern MS [14] is the key for successful proteomics analysis of ubiquitinated proteins [15, 16]. Ubiquitinated proteins are often present in cells at low abundance, and the modification is transient because some of the ubiquitination events are coupled with proteasomal degradation and deubiquitinating reactions are highly active in cells. To overcome these problems, it is essential to pre-enrich ubiquitinated proteins by various affinity strategies and to inhibit the loss of Ub conjugates by DUB inhibitors. The samples are then analyzed by highly sensitive MS with femtomolar or

even sub-femtomolar sensitivity, which enables the identification and quantification of hundreds to thousands of proteins in complex protein mixtures. Moreover, as protein modifications change the mass of targeted residues, a method has been developed to detect ubiquitination sites in modified proteins [17]. In addition, MS is an invaluable tool to characterize other protein components (e.g., E3 ligases and UBPs) involved in the Ub pathway. The similar strategies have been applied to analyze proteins modified by ULPs as well. To date, the technologies in yeast are more mature than those in other organisms. Interestingly, one of the earliest reports on the enrichment of Ub-conjugates was examined in *A. thaliana* expressing His-tagged Ub or its variant UbK48R [18]. Global analyses of ubiquitinated proteome in mammals or plants are anticipated to become feasible in the near future.

30.3 SPECIFIC METHODOLOGIES AND STRATEGIES

Many affinity purification schemes (Figure 30.2) have been developed to isolate ubiquitinated proteins from total cell lysates under native and denaturing conditions, including Ub-antibodies, UBPs, and epitope-tagged Ub (e.g., FLAG, HA-tag, *myc*-tag, His-tag, and biotin) [16]. The most successful approach is to purify ubiquitinated substrates using His- and/or biotin-tagged Ub under denaturing conditions, which reduces

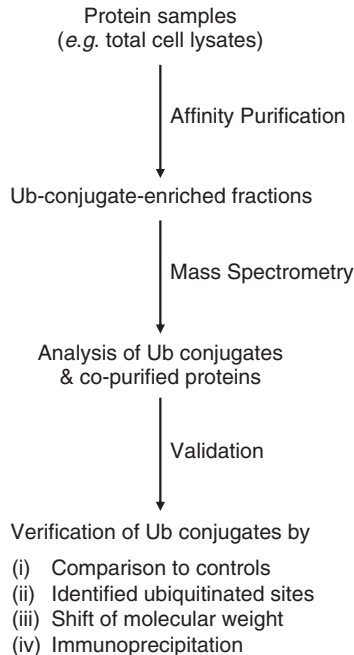


FIGURE 30.2. The scheme for analyzing ubiquitinated proteins. The Ub conjugates in complex protein mixtures are isolation by affinity methods, identified and/or quantified by MS and then confirmed by several strategies listed.

nonspecific protein binding to a minimal level and inhibits the activity of DUBs to prevent Ub conjugates from disassembly. Although the genetically tagging approach is applicable to yeast, cell cultures, and even transgenic animals, the samples without genetic manipulation (e.g., clinical specimen) cannot be analyzed unless nontagging methods are developed. Ub antibodies and UBPs may be used in such cases under native conditions or semi-denaturing conditions (e.g., 0.1% SDS) in the presence of DUB inhibitors [19]. During the purification, it is important to optimize experimental conditions, including the selection of column size and resins, the loading capacity of the column, and conditions for loading, washing, and eluting. While these parameters are often adjusted to balance the purity and yield of Ub conjugates, the sample purity is more critical with respect to decreasing co-purified contaminants. Considering that it is difficult to fully inactivate DUBs even in the presence of 8 M urea, it is better to perform the purification experiments rapidly, freeze the purified samples immediately, and avoid multiple freeze–thaw cycles.

Purified ubiquitinated proteins are commonly analyzed by two proteomics platforms: 1D SDS-PAGE followed by LC–MS/MS (GeLC–MS/MS), or LC/LC–MS/MS. In the former method, the protein mix is separated on a SDS gel and then subjected to in-gel digestion. The resulting peptides are resolved on a RP column coupled with MS. This approach permits the acquisition of apparent MW information of identified proteins, because the proteins are excised from gel lanes according to the molecular marker. The MW information greatly facilitates the separation of real Ub conjugates from contaminants, because ubiquitination leads to dramatic mass shift of target proteins. In the protocol of LC/LC–MS/MS, Ub-conjugate-enriched samples are directly digested by trypsin in-solution followed by prefractionation with strong cation exchange (SCX) chromatography. Every SCX fraction is further separated by RP–LC–MS/MS. In contrast to GeLC–MS/MS, direct in-solution digestion of Ub conjugates in a mixture provides better sensitivity for detection, because a single protein may be modified by Ub to form a ladder of species with mono- or poly-Ub chains, but they generate the same set of proteolytic fragments.

After protein identification by MS, it is important to validate whether the proteins are genuine Ub conjugates or co-isolated contaminants. While purification protocols can be optimized to reduce background proteins to a very low level, it is virtually impossible to remove all co-purified contaminants. Highly sensitive MS may be able to detect such contaminants. Several strategies have been developed to address this problem (Figure 30.2).

1. Remove proteins identified from mock purification. For example, during the purification of His-tagged Ub conjugates by Ni columns, a mock experiment could use cell lysate without His-tagged Ub, which permits the identification of common contaminants to be native His-rich and/or highly abundant proteins [17].
2. Confirm ubiquitinated proteins by identified modification sites. In MS analysis, trypsin cleaves conjugated Ub to a small dipeptide (–GG) tag that adds monoisotopic mass of 114.043 Da on modified Lys residues (Figure 30.3), and sometimes miscleavage in Ub generates a longer tag (–LRGG) [16]. Moreover,

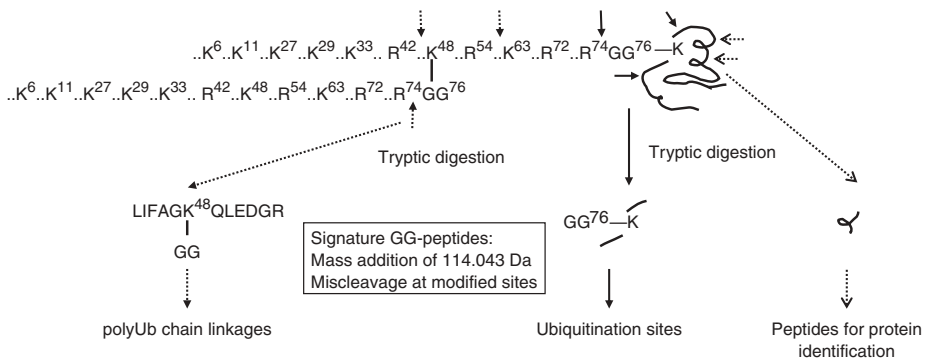


FIGURE 30.3. Analysis of a ubiquitinated protein by trypsin digestion and MS. Trypsin is a specific endoprotease that cleaves the peptide bonds at the C-termini of Arg or Lys unless followed by a Pro residue. When a Ub conjugate is digested with trypsin, the resulting peptides (on the right-hand side, dashed lines) can be used for determining the identity of the sequence. The peptide containing –GG tag (in the middle, solid lines) is generated by three cleavage events marked by arrows and is used for identifying the ubiquitination site by MS/MS. In addition, the –GG peptide digested from the polyUb chain (on the left-hand side, dashed lines) is useful for examining the chain linkage.

the –GG peptides contain a missed tryptic cleavage at the modified Lys sites. Based on these features, the Ub modification sites can be identified by unique MS/MS spectra of the –GG peptides during database searching. However, some ULPs, such as Nedd8 and ISG15, generate the same –GG peptides after trypsin digestion, and therefore it is prerequisite to separate those proteins by prefractionation. Another caveat is the difficulty to map all ubiquitinated sites in isolated Ub conjugates by current “shotgun” MS, because of undersampling (i.e., only a fraction of peptides are analyzed by MS/MS when many peptides are co-eluted together), lack of sensitivity (i.e., the recovery and/or ionization efficiency of some peptides are too low to be detected), and incompatibility (i.e., –GG peptides are too long or too short for the used proteomics platform; or the quality of MS/MS spectra is not good enough for effective database matching and quality filtering).

3. Evaluate the status of ubiquitination by the shift of apparent MW. In GeLC–MS/MS analysis, the MW information can be obtained from the SDS gel, which can be used to derive a virtual Western blotting result for identified proteins. The MW of unmodified proteins can be calculated from their amino acid sequence or even measured by GeLC–MS/MS. The comparison between native forms and ubiquitination forms would verify the status of Ub modification and lead to estimation of the length of Ub chains (~8 kDa for Ub monomer, ~16 kDa for di-Ub or double monoUb on two modification sites, and so on).
4. Affirm Ub conjugates of interest individually by immunoprecipitation. This independent approach has low throughput and can be used for analyzing selected proteins. The protein (or tagged version) is normally immunoisolated and probed

with antibodies against the protein itself and Ub. Ubiquitination is often verified by the detection of mass shift and HMW smear. Highly stringent condition (e.g., SDS) should be included to minimize co-immunoprecipitation of other ubiquitinated proteins. It should be cautious to interpret the result if the protein is overexpressed in cells.

30.4 EXPERIMENTAL RESULTS AND APPLICATIONS

MS Analysis of the Proteome Modified by Ub

A number of large-scale analyses of ubiquitinated proteins were performed, and they are different from each other with respect to starting samples, purification schemes, and MS technologies. In *S. cerevisiae*, Peng and co-workers [17, 20] described an extensive study of identifying ubiquitinated proteins from a mutated strain, in which all native Ub genes were deleted in the genome, and only His-tagged Ub was expressed from an introduced plasmid. His-tagged Ub conjugates were purified by Ni affinity chromatography followed by in-solution digestion and LC/LC-MS/MS. The mock experiment was carried out with an isogenic yeast strain expressing wild-type Ub. After removing 48 proteins in the mock, more than 1000 proteins were determined by MS in the pre-enriched sample. In addition, 110 ubiquitinated sites were defined in the sequences of 72 identified proteins, confirming the modification on these proteins. However, the majority of -GG peptides were missed in the analysis, highlighting the undersampling of “shotgun” MS. Although the exact FP rate is not known, it is believed that most of the identified 1000 proteins are modified by Ub [20]. Moreover, immunoprecipitation was used to assess the ubiquitination of 19 proteins, and none were found to be FP. Later, a TAP strategy under denaturing condition was developed to increase the purity of Ub conjugates, leading to identification of more than 250 proteins in yeast [21]. The tandem tag is composed of six His residues and a biotin motif. The tag is efficiently biotinylated by endogenous biotin ligases *in vivo*, which exist in both prokaryotic and eukaryotic cells. The extremely strong binding of biotin to streptavidin resin allows very stringent wash under denaturing conditions. In addition, Mayor et al. [22] implemented an affinity purification step by a Ub receptor Rpn10 in yeast prior to denaturing Ni column. The method permits the analysis of some ubiquitinated substrates that are recognized by a specific Ub receptor, which has potential for investigating Ub receptor-substrate interactions in cells.

Similarly, the ubiquitinated proteins from mammalian cells were examined by several groups with some success. Kirkpatrick et al. [23] found ~20 proteins in Ub-conjugate enriched samples from cells transfected with His-tagged Ub. Some researchers tested affinity purification of native Ub conjugates using Ub antibodies [19, 24]. Matsumoto et al. [19] attempted both native and denaturing conditions for the antibody affinity capture, and identified 670 and 345 proteins, respectively, indicating that ~50% of proteins enriched under native conditions are associated with the columns despite that they are not modified by Ub. In addition, Gururaja et al. [25] used an *in vitro* ubiquitination assay to synthesize His-tagged Ub conjugates from HeLa

cell extract for subsequent proteomics analysis. Taken together, the methodology in mammal at current stage is not as mature as that in yeast, because the mammalian cells contain much more native His-rich proteins for Ni affinity chromatography, and the expression level of tagged Ub is often lower than that of endogenous Ub in mammals. Further improvement of the method is required for routine analysis of proteins modified by Ub in more complex proteome. The tandem affinity tag [21] may be better suited for such a challenge.

The development of technology for profiling Ub conjugates makes it possible to compare ubiquitinated proteins under different physiopathological conditions. For instance, Hitchcock et al. [26] used the strategy to identify numerous membrane-associated Ub substrates in the ER-associated degradation (ERAD) pathway. In the experiments, they generated yeast strains with mutations in both the ER resident ubiquitinating enzyme UBC7 and the ERAD-associated protein NPL4. A total of 83 potential substrates in the ERAD pathway were found when compared to the wild-type strain.

Analysis of the Structure of PolyUb Chain

The consequence of ubiquitination on substrates may be highly dependent on the length and linkage of the attached Ub chains [15, 27]. For instance, a canonical K48-polyUb chain is the principal signal for targeting substrates for degradation by 26S proteasome, and efficient targeting requires a chain of at least four Ub molecules. Genetic and biochemical evidence support K29-linked chain as required for efficient degradation of substrates through the UFD (Ub-fusion degradation) pathway [28, 29]. Interestingly, the polyUb chains assembled on the UFD substrate are heteropolymers, which are initiated at K29 but elongated through K48 [29]. By contrast, polyUb chains linked through K63 has a nonproteolytic role in a wide range of other processes, and the physiological roles of the remaining linkages are poorly characterized.

The technology for identifying ubiquitinated sites allows direct measurement of Lys residues of ubiquitin that contribute to elongation of polyUb chains (Figure 30.3). Several large-scale analyses discovered unexpected complexity of polyUb assembly [17, 21]. In *S. cerevisiae*, all seven Lys residues in Ub were found to form Ub–Ub linkages [17]. Although an N-terminal-linked linear polyUb chain was not found in this analysis, a recent study reported that a mammalian Ub ligase can assemble this head-to-tail type of linkage *in vitro* [30]. Several proposed models for polyUb chain assembly were reviewed [31]. Quantitative analysis based on stable isotope-labeled internal standard revealed that the abundance ratio of K6:K11:K27:K29:K33:K48:K63 linkages to be ~40:120:20:15:5:100:40 during the log growth phase [32]. The total amount of non-K48 linkages is more than double the canonical K48 linkage. Moreover, increasing the endogenous concentration of Ub raised the level of almost all linkages, suggesting that the Ub level may be a rate-limiting step in the reaction. Furthermore, blocking proteasome activity by an inhibitor resulted in the increase of only K48- and K29-linked chains, recapitulating their roles in proteolysis and suggesting non-proteolytic functions of other linkages *in vivo* [32]. Another interesting finding is that forked polyUb chains (Figure 30.1) are present in cells, since several peptides with two –GG sites simultaneously were identified (K29–GG and K33–GG; K11–GG

and K27–GG; K27–GG and K29–GG) [17, 21]. In mammalian cells, K27-linked linkage was suggested by *in vitro* ubiquitination assay using a series of Ub mutants [33]. All other linkages were detected directly in multiple assays by MS [19, 32, 34]. More recently, Gygi's group has measured the frequency of different linkages on EGF receptor [34] and cyclin B1 [35]. The EGF receptor was found to be ubiquitinated with different lengths and linkages including about 49% mono-Ub, 40% K63, 6% K48, 3% K11, and 2% K29 linkages, consistent with the role of ubiquitination in protein trafficking and lysosomal degradation. *In vitro* ubiquitination of cyclin B1 generated mono-Ub and short polyUb chains linked by K48, K11, and K63, which can be degraded by proteasome effectively. More studies are needed to address if this phenomenon also occur *in vivo*.

Analysis of Other Components in the Ub Pathway

In addition to Ub substrates, proteomics approaches have been used to dissect other important players in the Ub pathway, including a variety of families of E3 ligases, DUBs and UBPs. Because the families of those proteins are becoming considerably large, it is not possible to list all relevant analyses in an exhaustive manner. This section will focus on the common proteomics methods that were applied in the fields and the introduction of some prehensive examples.

Identification of specific substrates of an E3 ligase is one of the most important steps to investigate its physiological function. One simple method is to search for defined ligase recognition motifs in a protein database. In the case of anaphase promoting complex (APC), which controls cell cycle transitions, it can interact and modify target proteins containing a destruction box motif (RxxLxxxxN) and/or a KEN box [36, 37]. However, the bioinformatics method is usually associated with substantial FPs and FNs. Considering that E3 ligases bind their substrates with certain affinity, it is possible to utilize the ligases to capture potential substrates, and vice versa. For example, Angers et al. [38] used a tandem-affinity (streptavidin- and calmodulin-binding motifs) tagged substrate to pull down the cullin-3 ligase complex. To further differentiate nonspecific binding proteins from real substrates after isolation of protein complexes, an *in vitro* ubiquitination assay can be implemented followed by separation of ubiquitinated species. Sato et al. [39] used this three-stage method to identify many substrates of the BRCA1-BARD1 Ub ligase. Alternatively, Kee et al. [40] used GST-fusion ligase Rsp5 to isolate substrates from pre-enriched Ub conjugates in yeast. In addition, recombinant protein substrates can be screened during *in vitro* high-throughput reconstituted reactions by specific ligases [41].

The DUBs consist of at least five families: Ub C-terminal hydrolases (UCHs), Ub-specific proteases (USPs), MJD proteases, OTU proteins, and JAMM motif metalloproteases [10]. The first four families are cysteine proteases, and JAMMs are zinc-dependent metalloproteases. Activity-dependent purification coupled with MS has been used to identify and characterize DUB enzymes in various cells and tissues. In the method developed by Borodovsky et al. [42], HA-tagged Ub derivatives were made to contain suicide thiol-reactive groups at its carboxyl termini, which react with the Cys active sites of DUBs in cell lysate. The resulting proteins were affinity-captured

and analyzed by MS. Besides the identification of at least 20 known DUBs, they discovered another new class of DUBs with OTU domains. Moreover, Ovaa et al. [43] used this method to profile active DUBs in normal, Epstein–Barr virus-infected, and malignant human cells, revealing some functional relationships between specific DUBs and virus infection or the induction of mitogen activity.

UBPs contain one or more UBDs that are structurally diverse motifs, composed of 20–150 amino acid residues [12]. UBDs can be found in Ub enzymes and Ub receptors that recognize mono-Ub or polyUb chains on modified substrates (Figure 30.1). A large number of Ub receptors (>250 in human) were identified by the methods of Y2H analysis, bioinformatics prediction and Ub-affinity isolation. Using Y2H assay, Bienko et al. [44] designed a mutant Ub as bait to capture two novel UBDs (UBM and UBZ) in all Y-family TLS polymerases and demonstrated their important regulatory functions. Although the affinity of UBDs to Ub is generally low (K_d within 2–750 μM) [12], the *in vitro* affinity purification is effective to enrich some UBPs. For instance, Russell and Wilkinson [45] developed a synthesized nonhydrolyzable analogue of K29-linked polyUb chain as affinity matrix to identify two binding partners, isopeptidase T and Ufd3, from yeast lysate.

Analysis of the UBL Pathways

The methods developed for the analysis of Ub pathway can be readily adapted to study the modifications by ULPs. SUMO is the most characterized ULP that plays a regulatory role in many cellular processes [13]. After the first large-scale analysis of ubiquitinated proteome [17], over 10 studies were performed to search for sumoylated proteins in different model systems. The details on affinity tags and isolation strategies were summarized in a recent review paper [32]. Among those, Wohlschlegel et al. [46] performed one of the most comprehensive analyses in yeast based on His-tag and identified 271 putative SUMO modified substrates. Like Ub, the precaution of co-purified proteins should be taken into consideration and follow-up validation of the sumoylation is needed. Different from protein ubiquitination, sumoylated Lys sites may be found in a consensus motif $\phi\text{KxD/E}$, where ϕ is a hydrophobic residue (most commonly leucine, isoleucine and valine) and x is any amino acid [47]. Moreover, MS/MS allows the determination of sumoylated sites. Tryptic digestion of SUMO generates much longer tag than –GG on substrates (e.g., EQIGG from yeast SUMO ortholog SMT3; a 19-aa-long tag from human SUMO-1). The 19-aa long tag forms complex fragmentation pattern that can be interpreted by specifically designed database-matching algorithm [48]. An alternative method is to mutant SUMO-1 sequence to shorten the tag to GG [49]. In addition, similar proteomics strategies were used to investigate other ULPs, such as ISG15 [50] and Nedd8 [51].

30.5 CONCLUSIONS

The analytical tools for protein ubiquitination have been quickly established in recent years in parallel with the development of modern MS. These high-throughput methods

have expanded the fields of proteins modified by different members in the Ub superfamily. It is believed that thousands of proteins in eukaryotic cells are regulated by protein ubiquitination. MS analyses have also revealed the surprisingly complexity of polyUb topologies that might provide difficult signals for the downstream events. The analysis of other components in ubiquitination, such as E3 ligase complexes, DUBs, and Ub receptors, are benefited from the advance of proteomics as well.

30.6 FIVE-YEAR VIEWPOINT

In the near future, the refinements of proteomics methods are anticipated, including the improvement of efficiency of affinity capture, higher detection sensitivity and throughput from more powerful mass spectrometers, and the seamless combination with quantitative approaches based on label-free methods or isotope-labeling strategies. The progress will largely reduce FPs and FNs in current proteomics approaches. More recently, monitoring the dynamics of >2000 phosphorylated proteins during a time course has become a reality [52], and the maturation of MS tools for ubiquitination in 5 years may allow us to profile as many as 2000 ubiquitinated species in different cellular conditions and genetic backgrounds.

ACKNOWLEDGMENT

This work was supported in part by NIH grants DK069580 and AG025688 to JP.

Editors Note. For further reading on the subject, readers are referred to Maor et al. [53] for a first view of ubiquitinated proteome in plants. Although Ub-binding proteins are abbreviated as UBP in the chapter for convenience, it should be noted that UBP is commonly used to refer Ub-specific proteases.

REFERENCES

1. Schlesinger DH and Goldstein G. 1975. *Nature* **255**:42304.
2. Ciechanover A, et al. 1980. *Proc Natl Acad Sci USA* **77**:1365–1368.
3. Hershko A, et al. 1980. *Proc Natl Acad Sci USA* **77**:1783–1786.
4. Ciechanover A, Finley D, and Varshavsky A. 1984. *Cell* **37**:57–66.
5. Finley D, Ciechanover A, and Varshavsky A. 1984. *Cell* **37**:43–55.
6. Hershko A and Ciechanover A. 1998. *Annu Rev Biochem* **67**:425–479.
7. Varshavsky A. 2005. *Trends Biochem Sci* **30**:283–286.
8. Ciechanover A and Ben-Saadon R. 2004. *Trends Cell Biol* **14**:103–106.
9. Cadwell K and Coscoy L. 2005. *Science* **309**:127–130.
10. Nijman SM, et al. 2005. *Cell* **123**:773–786.
11. Semple CA. 2003. *Genome Res* **13**:1389–1394.
12. Hicke L, Schubert HL, and Hill CP. 2005. *Nat Rev Mol Cell Biol* **6**:610–621.

13. Kerscher O, Felberbaum R, and Hochstrasser M. 2006. *Annu Rev Cell Dev Biol* **22**:159–180.
14. Aebersold R and Mann M. 2003. *Nature* **422**:198–207.
15. Kirkpatrick DS, Denison C, and Gygi SP. 2005. *Nat Cell Biol* **7**:750–757.
16. Xu P and Peng J. 2006. *Biochim Biophys Acta* **1764**:1940–1947.
17. Peng J, et al. 2003. *Nat Biotechnol* **21**:921–926.
18. Ling R, et al. 2000. *Anal Biochem* **282**:54–64.
19. Matsumoto M, et al. 2005. *Proteomics* **5**:4145–4151.
20. Peng J and Cheng D. 2005. *Methods Enzymol* **399**:367–381.
21. Tagwerker C, et al. 2006. *Mol Cell Proteomics* **5**:737–748.
22. Mayor T, et al. 2005. *Mol Cell Proteomics* **4**:741–751.
23. Kirkpatrick DS, et al. 2005. *Proteomics* **5**:2104–2111.
24. Vasilescu J, et al. 2005. *J Proteome Res* **4**:2192–2200.
25. Gururaja T, et al. 2003. *J Proteome Res* **2**:394–404.
26. Hitchcock AL, et al. 2003. *Proc Natl Acad Sci USA* **100**:12735–12740.
27. Pickart CM and Fushman D. 2004. *Curr Opin Chem Biol* **8**:610–616.
28. Johnson ES, et al. 1995. *J Biol Chem* **270**:17442–17456.
29. Koegl M, et al. 1999. *Cell* **96**:635–644.
30. Kirisako T, et al. 2006. *EMBO J* **25**:4877–4887.
31. Hochstrasser M. 2006. *Cell* **124**:27–34.
32. Xu P, et al. 2006. *Israel J Chem* **46**:171–182.
33. Hatakeyama S, et al. 2001. *J Biol Chem* **276**:33111–33120.
34. Huang F, et al. 2006. *Mol Cell* **21**:737–748.
35. Kirkpatrick DS, et al. 2006. *Nat Cell Biol* **8**:700–710.
36. King RW, Glotzer M, and Kirschner MW. 1996. *Mol Biol Cell* **7**:1343–1357.
37. Pflieger CM and Kirschner MW. 2000. *Genes Dev* **14**:655–665.
38. Angers S, et al. 2006. *Nat Cell Biol* **8**:348–357.
39. Sato K, et al. 2004. *J Biol Chem* **279**:30919–30922.
40. Kee Y, et al. 2006. *J Biol Chem* **281**:36724–36731.
41. Kus B, et al. 2005. *J Biol Chem* **280**:29470–29478.
42. Borodovsky A, et al. 2002. *Chem Biol* **9**:1149–1159.
43. Ovaa H, et al. 2004. *Proc Natl Acad Sci USA* **101**:2253–2258.
44. Bienko M, et al. 2005. *Science* **310**:1821–1824.
45. Russell NS and Wilkinson KD. 2004. *Biochemistry* **43**:4844–4854.
46. Wohlschlegel JA, et al. 2004. *J Biol Chem* **270**:45662–45668.
47. Rodriguez MS, Dargemont C and Hay RT. 2001. *J Biol Chem* **276**:12654–12659.
48. Pedrioli PG, et al. 2006. *Nat Methods* **3**:533–539.
49. Knuesel M, et al. 2005. *Mol Cell Proteomics* **4**:1626–1636.
50. Zhao C, et al. 2005. *Proc Natl Acad Sci USA* **102**:10200–10205.
51. Li T, et al. 2006. *Arch Biochem Biophys* **453**:70–74.
52. Olsen JV, et al. 2006. *Cell* **127**:635–648.
53. Maor R, et al. 2007. *Mol Cell Proteomics* **6**:601–610.

ANALYSIS OF THE N-GLYCOSYLATION OF PROTEINS IN PLANTS

Willy Morelle

31.1 INTRODUCTION

N-Glycosylation, in which the glycan is attached to specific asparagine residues, is considered to be one of the most important PTMs. *N*-Glycans have been found to participate in many biological processes. Each glycosylated site may contain many different glycan structures leading to pronounced heterogeneity (microheterogeneity). In addition, different sites may be only partially glycosylated (macroheterogeneity). Therefore, glycoproteins usually exist as complex mixtures of glycosylated variants (glycoforms), and full characterization of glycoproteins remains one of the most challenging tasks given to biochemists and bioanalysts.

Analysis of protein glycosylation can be performed at different levels of details. A glycomic analysis will give a profile of the total glycan population of a given glycoprotein or glycoprotein mixture, whereas a glycoproteomic analysis will provide the glycan repertoire at individual glycosylation sites. Complete structural analysis of glycans requires the determination of the monosaccharide composition, monosaccharide sequence, sugar branching, interglycosidic linkages, and anomeric configurations. As a consequence, it is often necessary to use a full range of various analytical methodologies to achieve complete structural assignments. These techniques may include

GC/MS, liquid chromatography, and MS in combination with chemical or enzymatic degradation and derivatization methods.

MS methods using MALDI and ESI play a crucial role in the elucidation of protein glycosylation and help to integrate glycosylation analysis into proteomics [1]. In this chapter, we summarize recent advances in the analysis of plant glycoproteins and their glycan structures using MS. The techniques are briefly described together with methods for glycan released and purification to MS.

31.2 BRIEF BIBLIOGRAPHIC REVIEW

Protein *N*-Glycosylation

N-Glycan biosynthesis is very similar in plants and animals, with the exception of late Golgi maturation events that generate unique types of complex plant glycans. However, it is not clear at which stage the process starts to differ [2]. *N*-Glycan biosynthesis begins during co-translation in the ER by the transfer of a $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ precursor to asparagine residues present in the tripeptide consensus sequon Asn-X-Ser/Thr (where X can be any amino acid except proline) of a nascent polypeptide chain. Once transferred and while the glycoprotein is transported along the secretory pathway, this precursor is processed by stepwise trimming and stepwise addition of new monosaccharide residues. After trimming by α -glucosidases I and II and α -mannosidase, the $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ precursor yields an intermediate glycan $\text{Man}_5\text{GlcNAc}_2$. This glycan is modified by several glycosyltransferases such as *N*-acetylglucosaminyltransferase I and II, xylosyltransferase, fucosyltransferase, and galactosyltransferase. Therefore, glycans moieties result from the complex coordination of several glycosyltransferases and glycosidases. These reactions result in a variety of oligosaccharide structures.

The oligosaccharide structures can be divided into three main classes: high-mannose-type *N*-glycans, complex-type *N*-glycans, and paucimannosidic-type *N*-glycans (Figure 31.1). The high-mannose-type *N*-glycans ($\text{Man}_{5-9}\text{GlcNAc}_2$) are identical in plant and animals, whereas plant complex type *N*-glycans differ substantially by the presence on the core of a $\beta(1-2)$ -xylose residue linked to the β -mannose and an $\alpha(1-3)$ -fucose residue linked to the proximal GlcNAc residue. These residues can be both allergenic and immunogenic in most mammals [3]. The $\beta(1-2)$ -xylose and $\alpha(1-3)$ -fucose residues are added respectively in the medial Golgi cisterna and in the trans Golgi [4]. These complex type *N*-glycans are also characterized by the presence of terminal Lewis^a sequences $\text{Gal}\beta(1-3)[\text{Fuc}\alpha(1-4)]\text{GlcNAc}$ that are added in the *trans* Golgi and *trans*-Golgi network but lack sialic acid residues [5]. These complex type *N*-glycans are present on secreted glycoproteins and highly expressed on the cell surface, whereas the paucimannosidic-type *N*-glycans ($\text{Xyl}_{0-1}\text{Fuc}_{0-1}\text{Man}_{5-9}\text{GlcNAc}_2$) that result from the elimination of terminal residues of complex type *N*-glycans are often present in vacuolar glycoproteins and seed storage glycoproteins.

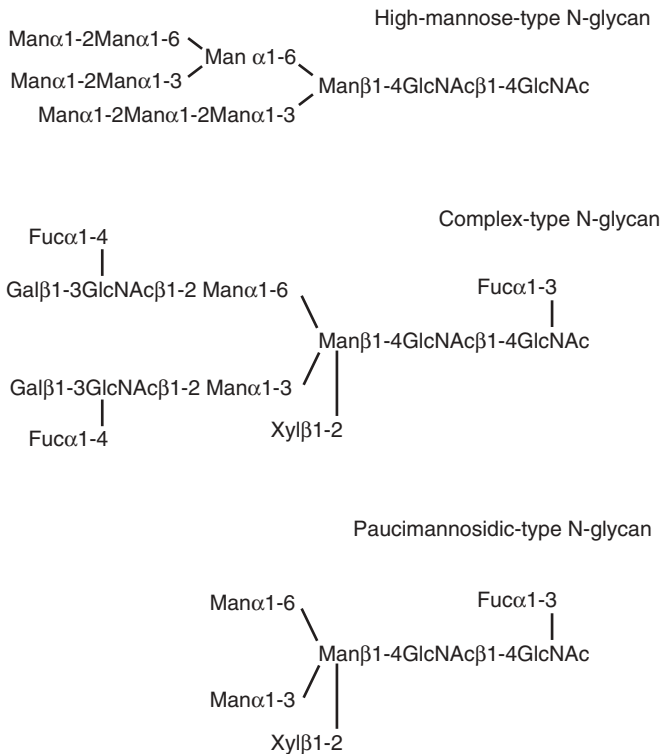


FIGURE 31.1. The plant N-linked glycans. This figure is designed to be indicative of some of the possible plant N-glycan structures.

Functions of N-Glycans

In plants, as in the other eukaryotes, the glycan chains in glycoproteins have been shown to play key functions in biological processes. These glycan chains also affect the physicochemical properties of a protein, such as its resistance to thermal denaturation, solubility stability, and protection from proteolytic degradation. The aid of N-glycans in protein folding and stabilizing the proteins is probably the single general function for glycoproteins. Useful tools to study the different roles of N-glycans are specific inhibitors of N-glycan biosynthesis and processing, site-directed mutagenesis of N-glycosylation sites, and mutants affected in the maturation of N-glycans.

A recurrent role of glycosylation is on protein folding in the ER. In the early steps of the protein biosynthesis, the presence of N-glycans is essential for the proper folding of the protein. Two chaperone proteins such as calnexin and calreticulin interact with the N-glycan chains, whereas a third chaperone protein, the BiP, recognizes the hydrophobic regions of the unfolded polypeptide [6, 7]. These chaperones interact

with monoglucosylated *N*-glycans of misfolded proteins; and the misfolded proteins remain confined to the ER, where they are degraded [8].

Several studies have demonstrated that in plants, initial *N*-glycosylation plays an essential role in growth and development [9, 10]. The α -glucosidase I mutants demonstrate that *N*-glycan trimming is essential for embryonic cell differentiation with respect to the accumulation of seed storage proteins, formation of protein bodies, cytokinesis and cellulose biosynthesis [9]. In contrast, XylT, FucT, or XylT/FucT double knockout lines were without a phenotype [11].

In mammals, phosphorylated high-mannose-type *N*-glycans is required for the transport of glycoproteins into lysosomes. In plants, the function of *N*-glycans in the targeting of proteins into the vacuole, the plant equivalent animal lysosome, has been extensively studied [12, 13]. These studies have shown that *N*-glycans do not participate in the targeting of glycoproteins to the plant vacuole. Conversely, it has been shown that *N*-glycans are necessary for transport and secretion of extracellular glycoproteins in plants [14, 15] but *N*-glycan processing is not required for glycoprotein secretion [16].

31.3 METHODOLOGY AND STRATEGY

Mass Spectrometry

Mass Spectrometry (MS), with the development of the MALDI and ESI methods in combination with modern separation methodologies, has become one of the most powerful and versatile techniques for glycosylation analysis. The determination of the native molecular mass and elucidation of the glycosylation sites, along with the macroheterogeneity (site occupancy) and microheterogeneity (complexity of glycoforms) of glycosylation, are important characteristics and can be determined using MS. Each of these ionization methods has its individual strengths and weaknesses (see Chapter 3). Because the MALDI experiment enables the predominant formation of singly charged quasimolecular ions and yields very few fragment ions, MALDI-MS is very often used as a first step in determining molecular mass of glycopeptides or glycans. The molecular masses are indicative of sugar compositions. ESI-MS also plays a crucial role in the structural analysis of glycosylation, providing not only relative molecular mass information but also sequence and branching information. Most ES instruments have tandem MS analyzers that allow the detection of fragment ions produced by CID of molecular ions in so-called MS-MS experiments. The parent ion is selected by the first mass analyzer for collision with an inert gas, and the fragment ions are screened by the second analyzer. In these experiments, detailed structural features of the glycans can be inferred from the analysis of the masses of the resulting fragments.

The development of these ionization methods catalyzed the development of new mass analyzers and complex multistage instruments. Q-TOF mass spectrometers provide CID spectra with significantly lower chemical noise than spectra acquired by triple Q or MALDI instruments, which have the advantages of high mass accuracy, great sensitivity, and resolution. These characteristics increase the dynamic range available which can be exploited to detect low-intensity interesting fragment ions

[17]. Another advantage of this type of instrument is the ability of the combination of liquid chromatography and MS–MS to separate, detect, and characterise glycans and glycopeptides. The recently MALDI–TOF–TOF-type instrument offers more options for glycosylation analysis. This mass spectrometer allows acquisition of high-quality fragment ion spectra from glycopeptides which provide information on the peptide sequence, the glycan attachment site, and the glycan structure [18]. Another instrument with great potential for glycosylation analysis is the FTICR–MS, which provides high mass resolution, mass accuracy, and MS/MS capabilities simultaneously. FTICR–MS offers several possible fragmentation techniques, such as ECD and infrared multiphoton dissociation (IRMPD), both of which have been employed to selectively investigate protein glycosylation. ECD is a soft fragmentation technique that cleaves the amino acid backbone without removing the PTM such as glycosylation [19], often resulting in near-complete sequence coverage and localization of the glycosylation modification. In contrast, IRMPD provide abundant fragment ions, primarily through dissociation at glycosidic linkages permitting structural information of the glycan to be acquired. Thus both fragmentation techniques yield complementary sequence information on glycopeptides with no release of the glycan. In addition, FTICR–MS/MS is sensitive and does not require extensive wet chemistry. These new instruments are responsible for the continuing progress in proteomics research and help to integrate glycosylation analysis into proteomics.

Identification of Glycoproteins

Glycan microheterogeneity may be responsible for glycoforms of the same protein appearing as multiple spots in 2D gels. This characteristic or an aberrant migration on 1-DGE can be used as a first indication that a protein is glycosylated. To confirm the presence of this PTM after 1- or 2-DGE, detection of glycoproteins can be achieved by several protein staining methods. The two most widely used methods for the detection of glycoproteins in gels and on blots involve reacting carbohydrate groups by a periodate/Schiff's base (PAS) mechanism and noncovalent binding of specific carbohydrate epitopes using lectin-based detection systems. The PAS method is based on the oxidation of carbohydrate groups to aldehydes using periodate. The carbohydrate groups can now be bound to a fluorescent substrate (Pro-Q Emerald dyes) or a tag (biotin hydrazide or digoxigenin hydrazide). Signal is detected directly using fluorescent substrates or indirectly using enzyme conjugates of streptavidin or antibodies to the tags. Gels and blots stained with Pro-Q EPS may be post-stained with SYPRO Ruby, allowing sequential two-color detection (green, red) of glycosylated and nonglycosylated proteins [20]. Lectins constitute a group of proteins with unique affinities toward carbohydrate structures. Numerous lectins with differing glycan specificities have been isolated from various sources and can be used to identify distinct types of carbohydrate structures on glycoproteins—for example, Con A from *C. ensiformis*, which binds to internal and nonreducing terminal α -mannosyl groups, detecting most glycoproteins [21]; and wheat germ agglutinin A (WGA), which recognizes *N*-acetyl-glucosamine and sialic acid residues. After its interaction with the blotted glycoprotein, the lectin is able to bind reporter enzymes such as HRP. These approaches relying on blot detection were well-described in a recent review by Fitchette et al. [22].

As recognition for their unique ability to bind glycans, lectin affinity chromatography (LAC) has now been developed for purification and concentration of glycoproteins and has become a powerful tool in glycoproteomics. Lectins can be used to reduce the complexity of the sample, to isolate glycoproteins, and to separate differently glycosylated isoforms. The most commonly used technique involves a small column containing lectin attached to agarose-based material [23], whereby the sample is loaded by using gravity-flow mode. After eliminating nonspecific binding, glycoproteins are eluted by displacement from the column with an elution buffer which has the composition of the loading buffer plus a haptene saccharide. LAC can use several lectins with broad specificity, such as Con A and WGA, to isolate the entire pools of glycoproteins present in the sample.

MALDI-TOF-MS and ESI-MS are powerful tools for the characterization of intact glycoproteins [24, 25]. To a variable extent, intact glycoproteins can be resolved to their individual glycoforms by both methodologies if the glycoproteins correspond to small proteins (up to 20–40 kDa) and/or contain a limited number of glycan chains. If the mass of the protein is known, the glycan mass can be determined by difference, and the numbers of constituent monosaccharides in terms of hexose, deoxyhexose, acetamidodeoxyhexose can be elucidated because plant *N*-glycans contain a limited number of type of monosaccharide residues. With larger molecules, resolution of glycoforms becomes increasingly difficult due to enormous heterogeneity arising from glycosylation macro- and microheterogeneity as well as other PTMs (sulfation, phosphorylation, hydroxylation, and carboxylations). Two strategies are then possible: (a) removal of the glycan or (b) cleavage of the peptide into smaller units. The measurement of protein MW before and after removal of the attached glycans provides information on the state of glycosylation, even though the individual glycoforms may not be resolved.

Analysis of Released *N*-Glycans

Most common MS approaches to characterize protein glycosylation are based on a chemical or enzymatic glycan release followed by labeling or derivatization, separation by chromatography or electrophoresis, and analysis in MS.

N-Glycans can be released from the glycoproteins using enzymatic or chemical methods. Chemical release has the advantage of being nonselective. Anhydrous hydrazine can release unreduced *N*-linked glycans from glycoproteins. This reagent cleaves peptide bonds including that between the *N*-linked glycan and asparagine. This chemical release suffers from several major disadvantages. As peptide bonds are destroyed, information relating to the protein is lost. Second, the acyl groups are cleaved from the *N*-acetyl amino sugars, calling for a reacetylation step. This reacetylation step can also add a small amount of *O*-acetyl substitution. Third, this reagent partially degrades *N*-glycans with an $\alpha(1-3)$ -Fuc residue linked to the proximal GlcNAc of the core [26]. For proteomic work, *N*-glycans are usually released using enzymatic methods. Several enzymes are commercially available. The most popular is PNGase F. PNGase F cleaves off the intact glycan as glycosylamine, which is readily converted to regular glycan. With few exceptions, PNGase F releases practically all protein-bound *N*-linked glycans, except those with fucose attached to the

3-position of the Asn-linked GlcNAc residue. Since these glycans are commonly found in plants and have been found to be sensitive to PNGase A, plant *N*-glycans must be released using PNGase A, which quantitatively releases both high-mannose and complex *N*-glycans in a nonspecific manner. PNGase A can also be used to release *N*-glycans from glycoproteins separated by 1-DGE [27] or 2-DGE [28].

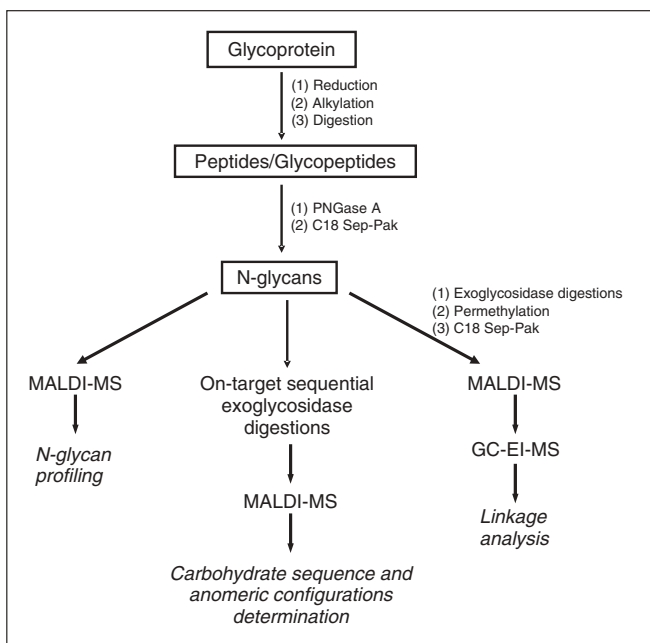
This enzymatic method makes it possible to recover reducing *N*-glycans, which can be derivatized via reductive amination with chromophores or fluorophores. Using this derivatization, Bardor et al. [29] developed a method for monitoring the *N*-glycosylation of plant glycoproteins by fluorophore-assisted carbohydrate electrophoresis (FACE). After enzymatic release from plant glycoproteins, *N*-glycans were reductively aminated to the fluorophore 8-aminonaphthalene-1,3,6-trisulfonic acid (ANTS) and separated according to their hydrodynamic size using high-resolution PAGE. This study demonstrated that electrophoretic mobilities of ANTS-labeled oligosaccharides are highly reproducible from gel to gel, allowing an easy identification of ANTS-labeled oligosaccharides isolated from plant glycoproteins. However, the analysis of complex mixtures of glycans cannot be performed in a single FACE analysis, and other techniques such as exoglycosidase digestions, chemical treatments, and analyses using MS are necessary to determine all *N*-glycan structures.

MALDI-TOF-MS is unique in its capacity to analyze complex mixtures, producing spectra uncomplicated by multiple charging or significant fragmentations. In addition, MALDI-TOF-MS provides significant tolerance against contaminations, such as salts or common buffers, as well as fast and easy sample preparation. MALDI is also more sensitive than the ESI technology. Besides, glycans yield intense signals corresponding to sodium cationized molecular species $[M+Na]^+$ in the positive ion mode, while multiple charged states are formed in ESI-MS that cause a loss in sensitivity as a result of splitting the signal of a single analyte. For these reasons, MALDI-TOF-MS is one of the most popular techniques for glycan analysis. The signal strength by N-linked glycans appears to reflect accurately the amount of material on the target, provided that the correct matrix is used [30]. Therefore, the relative quantities of glycans present in a mixture may be determined by MALDI glycan profiling.

The purity of the sample is a critical parameter for MS analysis of glycans. Compounds other than the analyte have an adverse effect on ion yield and should be removed prior the analysis. In addition, contaminants may induce signals, which complicate the spectra and make them difficult to interpret. The nature of the matrix and the method of sample preparation are of critical importance for obtaining strong signals from glycans. One of the most popular matrices for the analysis of glycans is 2,5-dihydroxybenzoic acid (DHB). The detection limit typically achieved with this matrix for analysis of neutral oligosaccharides is 50–100 femtomoles. MALDI-TOF-MS allows the structural elucidation of native and derivatized oligosaccharides. The derivatization of released oligosaccharides improves the ionization of oligosaccharides. Various derivatization strategies have been reported, including (a) permethylation and (b) reductive amination with aromatic or aliphatic amines. Despite the drawback of involving an additional wet chemistry step, permethylation derivatization offers several other advantages: As well as increasing the sensitivity of detection of molecular ions, this approach also leads to predictable fragmentations that give characteristic “maps”

of fragment ions at each amino sugar residue. In MS/MS experiments, the main fragmentation process of the permethylated glycans involves the cleavage of glycosidic bonds and gives rise to pertinent data on sequence and branching [31]. Recently, Viëtor et al. [32] have used this permethylation derivatization to investigate the structure of glycans *N*-linked to the proteins of the moss *P. patens* using MALDI-TOF-MS. Since the full structural analysis of a glycan includes the elucidation of all interglycosidic linkages and is usually performed using a methylation approach, MALDI-TOF-MS analysis and MS/MS experiments of permethylated glycans can represent the first step before a GC/MS analysis of chemically modified monosaccharides that are derived from methylation, hydrolysis, reduction, and acetylation of glycans (Box 31.1).

Several methods to prepare *N*-glycan samples with PNGase A can be used. The general strategy that we use for investigating the glycosylation pattern of plant glycoproteins is outlined in Box 31.1. In this method, to facilitate the release of *N*-glycans without resorting to detergent denaturation the purified glycoproteins or unfractionated extracts are first reduced and carboxamidomethylated. Reduced and carboxamidomethylated glycoproteins are then submitted to a proteolytic digestion to generate small peptides and glycopeptides before PNGase A digestion. PNGase A released glycans are separated from peptides using a Sep-Pak C18 cartridge. At this stage, PNGase A released glycans can be analyzed using MS before or after derivatization. If *N*-glycans are analyzed without derivatization, *N*-glycans need to be desalted on nonporous graphitized carbon solid-phase extraction cartridge. Otherwise, *N*-glycans can be directly permethylated and examined by MALDI-TOF-MS, nESI-Q-TOF-MS/MS, and linkage analysis after reverse-phase Sep-Pak C18 purification. Sep-Pak C18 is a simple, rapid method for separating permethylated glycans from remaining salts. If limited amounts of material are available, *N*-glycans are analyzed as mixtures. The composition of the molecular ions as deduced from their precise m/z values, when considered in conjunction with methylation analysis data, allow important structural conclusions to be drawn on picomolar amounts of components. The permethylated components observed in the MALDI-TOF-MS spectrum can be further analyzed using a nESI-Q-TOF-MS to assist sequence assignment [31]. This instrument is used to obtain fragmentation spectra in a short time with high mass accuracy which reveal detailed structural and in some cases linkage information on complex oligosaccharides. MS and MS/MS experiments rarely provide sufficient information to define carbohydrate structures rigorously. Hence, data from MS analyses of enzymatic degradations are usually required to supplement molecular and fragment ion information to define structural features such as sugar type, branching, and anomeric stereochemistry. Exoglycosidases remove monosaccharides from the nonreducing termini of glycan chains. After enzymatic treatment, the mixture is analyzed by MALDI-TOF-MS to determine the number of monosaccharides released by the enzyme. Using several well-defined enzymes in sequence, the molecular mass information after each digestion step thus reveals the sequence of the monosaccharide constituents of the glycan chains. Aliquots can also be taken after each digestion, permethylated, and examined by MALDI-TOF-MS and linkage analysis after reverse-phase Sep-Pak C18 purification. Comparison of linkage data before and after exoglycosidase treatment makes it possible to establish where the removed

Box 31.1

Strategy dedicated to the determination of the structures of the *N*-glycans of plant glycoproteins. This strategy can be applied to a purified glycoprotein or a mixture of glycoproteins. Reduced and carboxamidomethylated glycoproteins are submitted to a proteolytic digestion to generate small peptides and glycopeptides before PNGase A digestion. PNGase A released glycans are separated from peptides using a Sep-Pak C18 cartridge and desalted on nonporous graphitized carbon solid-phase extraction cartridge, PNGase A-released *N*-glycans are then characterized by MALDI–MS before and after on-plate sequential exoglycosidase digestions. Their methylated derivatives are also characterized by MALDI–MS before and after sequential exoglycosidase digestions and by linkage analysis. Structural assignments are based on molecular weight, susceptibility to exoglycosidase digestions, and linkage data.

monosaccharides were attached. Structural assignments of PNGase A-released glycans are based on MW, susceptibility to exoglycosidase digestions, and linkage data.

Analysis of Glycopeptides

Protein glycosylation should ideally be characterized at the glycopeptide level in order to obtain information about glycosylation site occupancy and heterogeneity. In a typical glycoproteomic experiment, the glycoprotein spot will be cut from the gel, destained, digested with a specific endoprotease, and analyzed using MS. This

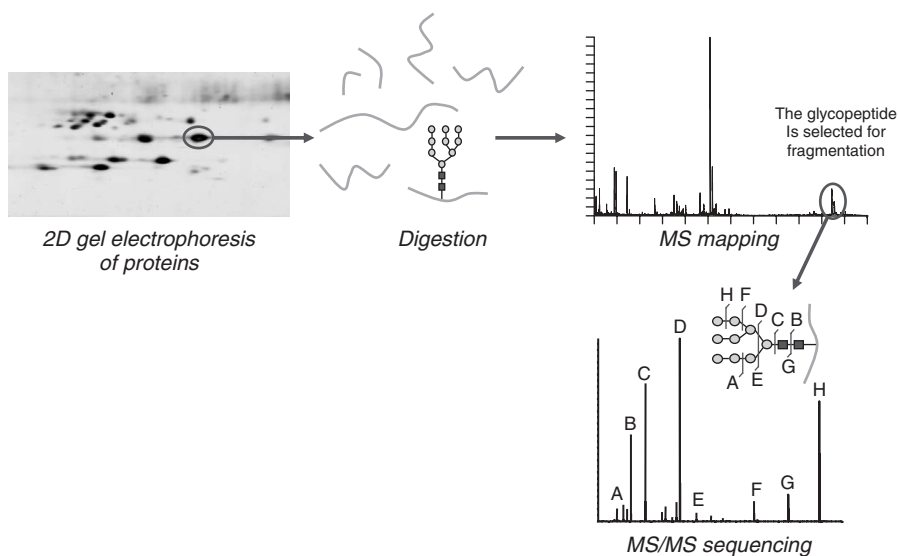


FIGURE 31.2. MS approach for the analysis of protein glycosylation. After separating the proteins by 2-DE, the resultant spots can be cut, destained, digested by trypsin, purified, and analyzed by MS. Glycopeptides are then selected for MS/MS sequencing.

typical glycoproteomic experiment is described in Figure 31.2. Tools that scan the different protein databases available allow the identification of the protein based on its *pI*, *M_r*, and the masses of the different peptides generated from the endoprotease digestion of the glycoprotein. Mass shifts resulting from the glycosylation of peptides can be detected, and structural information and characterization of the attached oligosaccharides can often be obtained in the form of putative structures based upon their apparent mass. To determine the actual glycan structure and glycosylation site, the digest is further subjected to fragmentation. MS/MS experiments are mostly performed using ES instruments with a Q-TOF analyzer, applying CID. However, the most frequently observed fragmentation is that corresponding to the cleavage of glycosidic linkages. Therefore, the fragmentation of glycosylated peptides, in most cases, provides information on the glycan moiety, but not on the attachment site.

It is rather difficult to identify glycopeptides using MS in a complex protein digest. Glycopeptide signals are often suppressed in the presence of other peptides. For this reason, the peptide and glycopeptide mixture can be separated by RP chromatography. Fractions are collected and analyzed using MS. Samyn-Petit et al. [33] used this strategy for the comparative analysis of the site-specific *N*-glycosylation of human lactoferrin produced in maize and tobacco plants. The different HPLC fractions were collected and analyzed by MALDI-TOF. Glycopeptides were then identified, based on theoretical masses of peptides from the reduced and carboxamidomethylated lactoferrin with various additional *N*-glycans. After deglycosylation, the peptides were also analyzed by nESI-Q-TOF-MS/MS to identify the glycosylation site. ESI-MS can also be used as an on-line HPLC detector for tryptic mapping of glycoproteins.

The major advantage of this technique is that carbohydrate-specific ions can be selectively detected at high sensitivity during the chromatographic separation of complex digest mixtures. Glycopeptide-marker ions in CID are LMW oxonium ions of m/z 163 (Hexose), 204 (HexNAc), and m/z 366 (Hexose-HexNAc), among others. These characteristic ions are produced by either increasing the orifice voltage in the ESI source or through CID within the collision cell of the triple-Q mass spectrometer. Although the lability of the glycosidic bond favors a sensitive and selective detection of glycopeptides, it affects the fragmentation of the peptide backbone negatively. The low relative abundance of the peptide fragment ions hinders their use for peptide sequence determination in *N*-glycosylated peptides analysis.

To overcome these problems, the potential of new mass spectrometers such as FTICR and MALDI-TOF-TOF has been explored for the analysis of *N*-glycosylated peptides [18, 19]. Two different types of cleavage are observed in the TOF-TOF fragmentation spectra: First, cleavages of peptide bonds yield fragments with no fragmentation of the glycan. Second, fragmentation of the glycan moiety is mostly characterized by cleavage of glycosidic bonds, thus revealing information on glycan structures. Therefore MALDI-TOF-TOF-MS/MS resembles FTICR-MS with tandem ECD, where peptide fragmentation occurs faster than the glycan loss, allowing the localization of the glycosylation sites. These instruments provide detailed sequence and site-of-attachment data for glycosylated proteins. Recently, Bykova et al. [34] used a QqTOF-type instrument with a MALDI ionization source to elucidate the glycans' structure and their sites of attachment in a plant protease. The authors characterized four *N*-glycosylation sites in tomato pathogenesis-induced subtilisin-like P69B protease with site-specific distribution of glycoforms, using a combination of 2-DGE, proteolytic digestion, LC, and MALDI-Qq-TOF-MS/MS. MS/MS spectra allowed sequencing and deduction of the composition for both the peptide and oligosaccharide parts of each glycoform in a single step. MALDI-Qq-TOF revealed heterogeneity at different levels, including different glycan side-chain modifications, and heterogeneity of oligosaccharide structures on the same glycosylation site. The authors showed that MALDI-QqTOF-MS/MS analysis coupled with HPLC separation of protein digest is a very sensitive and powerful technique with respect to identification of glycosylation sites from very small quantities of sample such as protein spots originating from 2D gels. Besides, detection of modifications of methionine on the peptide backbone of glycopeptide illustrates the potential of this approach as a valuable tool for the straightforward analysis of protein glycosylation with other types of modifications.

31.4 CONCLUSIONS

PTM is an important feature of a proteome, frequently conveying a specific biological activity or role to a protein. Glycosylation is one of the most common PTMs of proteins and is a target for proteomic research. In addition, elucidation of protein glycosylation is not only necessary to understand glycoprotein structures and functions in plants but is also crucial for biopharmaceutical production [35]. Glycosylation analysis remains a difficult task. There are fundamental differences between the identification of a protein and the analysis of its glycosylation. First, full sequence coverage

must be obtained in order not to miss the glycosylated amino acids, whereas partial sequence coverage is usually adequate for unambiguous protein identification. Second, carbohydrate primary structure is far more complex than peptide primary structure and hence requires more analysis steps, thus further increasing the demand for yet more sensitive analytical methods. Structural analysis of glycoproteins is currently undergoing rapid expansion owing to developments in MS instrumentation and associated techniques, which provide the largest amount of information, such as determination of glycosylation sites, macroheterogeneity (site occupancy), and microheterogeneity (complexity of glycoforms). MS can also give representative glycan profiles, and fragmentation of glycans provides most of the detail. Other techniques such as exoglycosidase digestions and chemical treatments are also necessary to supplement the mass data. Therefore modern glycoanalysis is a multimethodological task in which MS is playing a fundamental role, and the methods described here are very efficient to determine *N*-linked glycan structures and help to integrate glycosylation analysis into plant proteomics.

31.5 FIVE-YEAR VIEWPOINT

Plants have emerged as an alternative to current systems for the production of therapeutic proteins. Plants offer many potential benefits for the production of recombinant pharmaceutical proteins in terms of cost, scalability, and safety. A major advantage of plants is their ability to perform most of the PTMs that are required for the bioactivity and pharmacokinetics of recombinant therapeutic proteins. However, they are not able to reproduce human glycosylation perfectly. Several attractive strategies to humanize plant *N*-glycans have been proposed; in this context, fast analytical approaches are necessary for the analysis of protein glycosylation. Developments in MS are crucial for glycan analysis and should allow high-throughput structural analysis of *N*-glycans at the amounts available in plant proteomic studies. Fragmentation spectra can provide a great deal of information and in most cases should replace exoglycosidase digestions. Extensive databases of MS/MS spectra should be available, and high-throughput tools to help glycobiologists to interpret MS/MS spectra should be essential in glycomics and glycoproteomics. The exoglycosidase digestions will probably only be used to help elucidate unusual glycan structures. On the glycopeptide level, it is rather difficult to identify glycopeptides using MS in a complex protein digest. To overcome this problem, LAC is a very effective tool to target glycopeptides and can be used to remove efficiently non-glycosylated peptides and concentrate glycopeptides for proteomic purposes. The uses of lectin-based chemistries for high sensitivity glycomic/glycoproteomic structural analysis are likely to grow. Q-TOF, Q-FTICR mass spectrometers, and ion trap mass spectrometers using the complementary techniques of CID and ETD are particularly versatile instruments for glycopeptide analysis and should be very useful to integrate glycosylation analysis into plant proteomics.

REFERENCES

1. Morelle W, et al. 2006. *Proteomics* **6**:3993–4015.
2. Strasser R, et al. 2006. *Plant J* **45**:789–803.
3. Faye L, et al. 1993. *Anal Biochem* **209**:140–180.
4. Fitchette-Lainé AC, et al. 1994. *Plant J* **5**:673–682.
5. Séveno M, et al. 2004. *Nat Biotechnol* **22**:1351–1352.
6. Denecke J, et al. 1995. *Plant Cell* **7**:391–406.
7. Lupattelli F, et al. 1997. *Plant Cell* **9**:597–609.
8. Ivessa E. 1999. *Trends Cell Biol* **9**:130–131.
9. Boisson M, et al. 2001. *EMBO J* **20**:1010–1019.
10. Lerouxel O, et al. 2005. *Plant J* **42**:455–468.
11. Strasser R, et al. 2004. *FEBS Lett* **561**:132–136.
12. Bollini R, et al. 1985. *Physiol Plant* **65**:15–22.
13. Bednarek SY and Raikhel NV. 1991. *Plant Cell* **3**:1995–1206.
14. Hori H and Elbein AD. 1981. *Plant Physiol* **67**:882–886.
15. Driouich A, et al. 1989. *Planta* **180**:96–104.
16. Lerouge P, et al. 1996. *Plant J* **10**:713–719.
17. Morelle W and Michalski JC. 2004. *Electrophoresis* **25**:2144–2155.
18. Wührer M, Hokke CH and Deelder AM. 2004. *Rapid Commun Mass Spectrom* **18**:1741–1748.
19. Håkansson K, et al. 2001. *Anal Chem* **73**:4530–4536.
20. Steinberg TH, et al. 2001. *Proteomics* **1**:841–855.
21. Kornfeld K, Reitman M and Kornfeld R. 1981. *J Biol Chem* **256**:6633–6640.
22. Fitchette AC, et al. 2006. *Methods Mol Biol* **335**:317–342.
23. Cummings RD. 1994. *Methods Enzymol* **230**:66–86.
24. Harvey DJ. 2005. *Expert Rev Proteomics* **2**:87–101.
25. Morelle W and Michalski JC. 2005. *Curr Anal Chem* **1**:29–57.
26. Misaki R, et al. 2003. *Glycobiology* **13**:199–205.
27. Kolarich D and Altmann F. 2000. *Anal Biochem* **285**:64–75.
28. Balen B, et al. 2006. *J Proteome Res* **5**:1658–1666.
29. Bardor M, et al. 2000. *Electrophoresis* **21**:2550–2556.
30. Harvey DJ. 1993. *Rapid Commun Mass Spectrom* **7**:614–619.
31. Morelle W, Faïd V, and Michalski JC. 2004. *Rapid Commun Mass Spectrom* **18**:2451–2464.
32. Viëtor R, et al. 2003. *Planta* **218**:269–275.
33. Samyn-Petit B, et al. 2003. *Eur J Biochem* **270**:3235–3242.
34. Bykova NV, et al. 2006. *Anal Chem* **78**:1093–1103.
35. Bardor M, et al. 2006. *Curr Opin Struct Biol* **16**:576–583.

FUNCTIONAL ANALYSIS AND PHOSPHORYLATION SITE MAPPING OF LEUCINE-RICH REPEAT RECEPTOR-LIKE KINASES

Steven D. Clouse, Michael B. Goshe, Steven C. Huber, and Jia Li

32.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Biological Function of LRR RLKs

The completed genome sequences of *A. thaliana* and rice have revealed very large multigene families encoding predicted proteins with an organization of functional domains similar to that of animal receptor kinases, including a putative extracellular ligand-binding domain, a single-pass TM sequence, and an intracellular cytoplasmic kinase domain [1]. Such structural features suggest a role for these proteins in extracellular signal perception followed by intracellular transduction of the signal via phosphorylation of specific targets. More than 220 of these RLKs in *Arabidopsis* and nearly 400 in rice have an extracellular domain containing multiple LRRs and are thus classified as LRR RLKs. LRR RLKs can be further divided into 15 subfamilies based on sequence alignment [2]. A small subset of these LRR RLKs have been characterized in detail by genetic and biochemical analyses and have been shown to play essential roles in regulating plant growth and development as well as in defense responses to various pathogens [3].

BRI1 and BAK1, both critically involved in signal transduction for the essential plant hormone BL, have been particularly well-characterized [4–8] and share some mechanistic similarities to animal receptor kinase function, including ligand-dependent oligomerization, activation of the intracellular kinase domain by phosphorylation of specific Ser and Thr residues, and transduction of the signal downstream by phosphorylation of selective substrates [9]. To fully characterize plant LRR RLK function across the entire family, it is essential to understand the role of receptor dimerization and kinase domain autophosphorylation, including identification of specific phosphorylation sites and their functional significance. Moreover, identification of LRR RLK kinase domain substrates and cataloging structural requirements for substrate recognition and phosphorylation are also necessary for a complete picture of LRR RLK action. We are undertaking such a functional analysis of the LRR RLK family as part of the U.S. NSF *Arabidopsis* 2010 project (<http://www4.ncsu.edu/~sclouse/Clouse2010.htm>) using a multilevel functional analysis employing molecular genetics, proteomics, kinase biochemistry, and phosphoprotein analysis with MS.

Proteomics Approaches to LRR RLK Analysis

Successful proteomic analysis of LRR RLKs requires efficient methods for membrane isolation, solubilization, and fractionation in order to obtain sufficient purification and enrichment of LRR RLKs of possibly low abundance for detection by MS-based approaches. Multiple extractions of purified PMs from *Arabidopsis* cell suspension cultures followed by 1-DGE and in-gel tryptic digestion allowed the identification of 100 PM proteins using multiple MS analyses [10]. However, only two of the identified proteins were LRR RLKs. Subsequent studies in *Arabidopsis* using purified PM from suspension-cultured cells incorporated additional enrichment and labeling techniques for phosphopeptides, which dramatically increased the number of LRR RLKs identified [11]. While plant cell suspension cultures provide an abundant and uniform population of cells, their undifferentiated state makes them less desirable for studying LRR RLKs, which are involved in complex regulatory pathways associated with growth and development of whole tissues and organs. Relatively few studies of the *Arabidopsis* PM proteome have utilized whole plants. Alexandersson et al. [12] purified PM from *Arabidopsis* leaves and applied gradient SDS-PAGE, in-gel tryptic digestion, and LC-MS/MS analysis to identify 238 putative PM proteins, seven of which were LRR RLKs. Nelson et al. [13] used whole *Arabidopsis* seedlings grown in shaking liquid culture to isolate 31 RLKs from purified PM, 21 of which were identified as LRR RLKs.

For our studies, we chose light-grown whole *Arabidopsis* seedlings subjected to a variety of physiological conditions for functional proteomics studies of LRR RLKs. An efficient extraction protocol for PM proteins was developed and is being combined with a novel isotope-coded affinity procedure prior to 2D LC-MS/MS (see below). The goal is to identify a subset of 30 LRR RLKs that are phosphorylated in response to different hormone treatments and abiotic stresses and then develop a phosphorylation site database for these proteins. Individual phosphorylation sites are

being studied in detail by immunoprecipitation of specific LRR RLKs from transgenic lines expressing FLAG-tagged [14] versions of the protein, followed by LC–MS/MS analysis. Selected phosphorylation sites are then analyzed for functional significance using genetic and biochemical approaches that require a mutant LRR RLK with an observable phenotype. As a resource for this project and the *Arabidopsis* community in general, we have generated over a thousand different constructs for LRR RLK analysis in both transgenic plants and *E. coli*, using the Gateway vector system [15]. Tagged recombinant proteins for the entire LRR RLK family are being used for *in vitro* analysis of autophosphorylation activity, substrate preference, and pairwise interactions. Guided by *in vitro* interaction studies, *in planta* interactions of selected LRR RLKs are analyzed by co-immunoprecipitation of epitope-tagged proteins from transgenic plants and by overexpression of kinase inactive mutants to search for dominant negative phenotypes. As a prototype for this overall experimental approach, the LRR RLKs, BRI1, and BAK1 have been studied in detail.

BRI1/BAK1: A Model for Analysis of LRR RLK Heterodimerization, Phosphorylation, and Function

BRs are endogenous plant growth-promoting hormones that regulate cellular expansion, differentiation, and proliferation, and the phenotype of mutants affected in BR biosynthesis or signaling clearly show that these plant steroids are essential for normal plant development [16]. Extensive genetic and biochemical analyses strongly support the role of at least two LRR RLKs in active BR signaling. BRI1, the BR receptor, is an LRR RLK with 25 LRRs in the extracellular domain (Figure 32.1), including a 70-amino-acid “island” between LRR 21 and 22 [4]. Recent studies have demonstrated that this island domain along with LRR 22 is directly involved in binding the steroid ligand [17]. Null mutations in either the extracellular domain or cytoplasmic kinase domain of BRI1 lead to an extremely dwarf phenotype with multiple developmental defects observed in *Arabidopsis* and several crop species [18, 19]. The importance of heterodimerization in BR signaling was revealed by the identification of a second LRR RLK, BAK1, using two independent screens [5, 6]. BAK1 has a much smaller extracellular domain than BRI1 and lacks the 70-amino-acid island (Figure 32.1) and is thus not likely to bind BR directly. Overexpression of BAK1 suppresses weak *bri1* alleles but not null alleles, suggesting that at least partial BRI1 function is a prerequisite for BAK1 action.

Several studies have shown that BRI1 and BAK1 directly interact with each other both *in vitro* and *in vivo* [5, 6, 20]. This interaction appears to be BR-dependent, since depletion of endogenous BR by treatment of *Arabidopsis* seedlings with a BR biosynthetic inhibitor greatly reduced the affinity of these two RLKs, while supplementation with exogenous BR significantly strengthened the BRI1–BAK1 association [21]. Phosphorylation of the cytoplasmic domains of both BRI1 and BAK1 was also shown to be BR-dependent. Multiple *in vivo* phosphorylation sites for BRI1 and BAK1 have now been identified by immunoprecipitation followed by LC–MS/MS analysis. The functional significance of phosphorylation on specific Ser and Thr residues of BRI1 was assessed *in vitro* with respect to kinase activity, as well as *in vivo* by the ability of

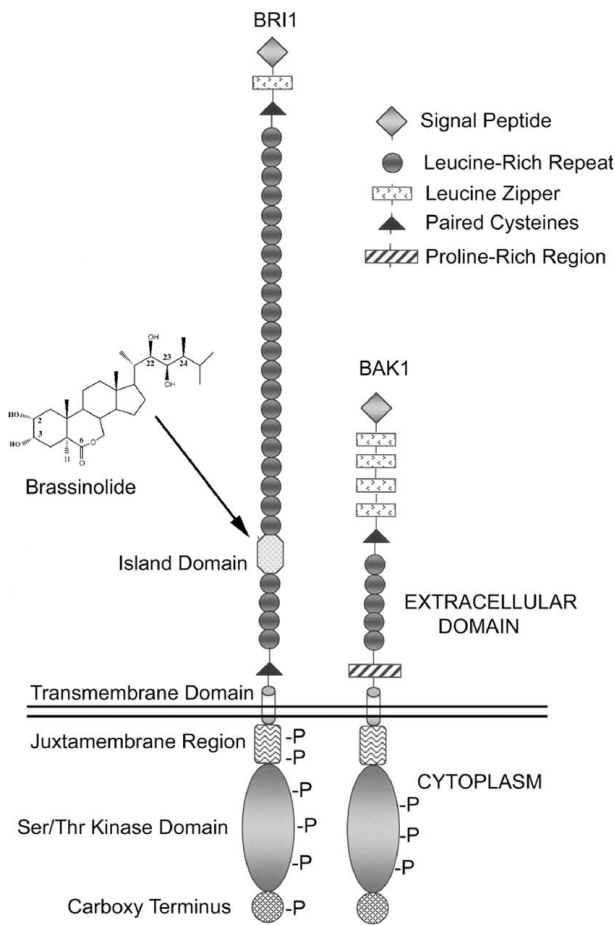


FIGURE 32.1. Structure of two LRR RLKs, BRI1, and BAK1, involved in BR signal transduction. Brassinolide, the plant steroid hormone, binds directly to a region of BRI1 containing the 70-amino-acid island and LRR 22. BRI1 and BAK1 form a heterodimer and are phosphorylated on multiple residues of the cytoplasmic domain. Both heterodimerization and phosphorylation are ligand-dependent.

mutated versions of a BRI1–FLAG transgene to rescue the dwarf phenotype of a BRI mutant [21]. Current models regarding early events of BR signaling suggest that BRI1 forms a homodimer upon BR binding, followed by activation of the BRI1 catalytic kinase domain via phosphorylation [19]. The activated BRI1 then interacts with and activates BAK1. Some features of these models suggest that BR signal transduction shares mechanistic similarities with animal RTKs and/or the TGF- β Ser/Thr receptor kinases. The extensive biochemical and genetic analysis of BRI1 and BAK1 make these two of the best understood LRR RLKs and provides a model for examination of mechanisms of other LRR RLK signaling pathways.

32.2 SPECIFIC METHODOLOGIES AND STRATEGIES

Plant Membrane Proteomics

As a foundation for our quantitative phosphoproteomic analysis of LRR RLK function, we developed a generally applicable method for plant membrane proteomics using a buffered methanol solution to facilitate miscible extraction, solubilization, tryptic digestion, and fractionation using strong CAX. The use of non-ionic detergents such as Triton X-100 and Brij-58 for membrane protein solubilization has been widely employed in numerous proteomic studies [22]. While effectively solubilizing many membrane proteins, such detergents may interfere with downstream LC–MS/MS analysis by affecting chromatographic separations and suppressing peptide ionization and MS detection [23]. The use of MS-compatible organic solvents such as methanol for membrane protein solubilization, followed by direct in-solution tryptic digestion and LC–MS/MS, was found to be highly effective in proteomic studies of bacterial and human membranes [23]. We compared this approach with non-ionic detergent solubilization of *Arabidopsis* microsomal fractions isolated from whole-plant tissue and found that while the total number of proteins obtained using the non-ionic detergent was higher than the total obtained by methanol, the percentage of membrane proteins obtained by the use of organic solvent was much greater. Of particular relevance to our project, the number of LRR RLKs identified with the methanol approach was three times that of the non-ionic detergent. The methanol method was also highly efficient when applied to purified PMs isolated by phase partitioning [24], which, as expected, increased the number of LRR RLKs identified. Details of the methanol method have been published [25]. This protocol is being combined with ICAT procedures to quantify phosphorylation state changes in LRR RLKs from plants grown under different conditions.

Isotope-Coded Affinity Approaches for Quantitative Analysis of Phosphorylation Sites

Isotope coding is the process by which proteins (or their corresponding peptides) are labeled with a combination of stable isotopes that are used to differentiate control from treated samples while permitting relative quantification of protein levels and phosphorylation events. One fractionation technique amenable to LC–MS/MS analysis using “bottom-up” proteomic approaches is IMAC for phosphopeptide enrichment. IMAC relies on the affinity of phosphate groups for metal ions such as Ga^{3+} or Fe^{3+} bound to chelating reagents tethered on solid-phase supports. Carboxylic acid side chains in nonphosphorylated peptides promote nonspecific binding to the IMAC resin, which can be reduced by converting these acidic groups to their corresponding methyl esters [26]. The esterification is amenable to stable isotope labeling with d_0/d_3 -methanol and can be used to perform comparative proteomic studies [27] (Figure 32.2A, C). The IMAC procedure has been successfully applied in a number of phosphoproteomics studies [11, 28] and is also being used in our LRR RLK proteomic analysis. Other isotope-based methods of potential applicability, including the AQUA method [29] and iTRAQ [30], are described elsewhere in this book.

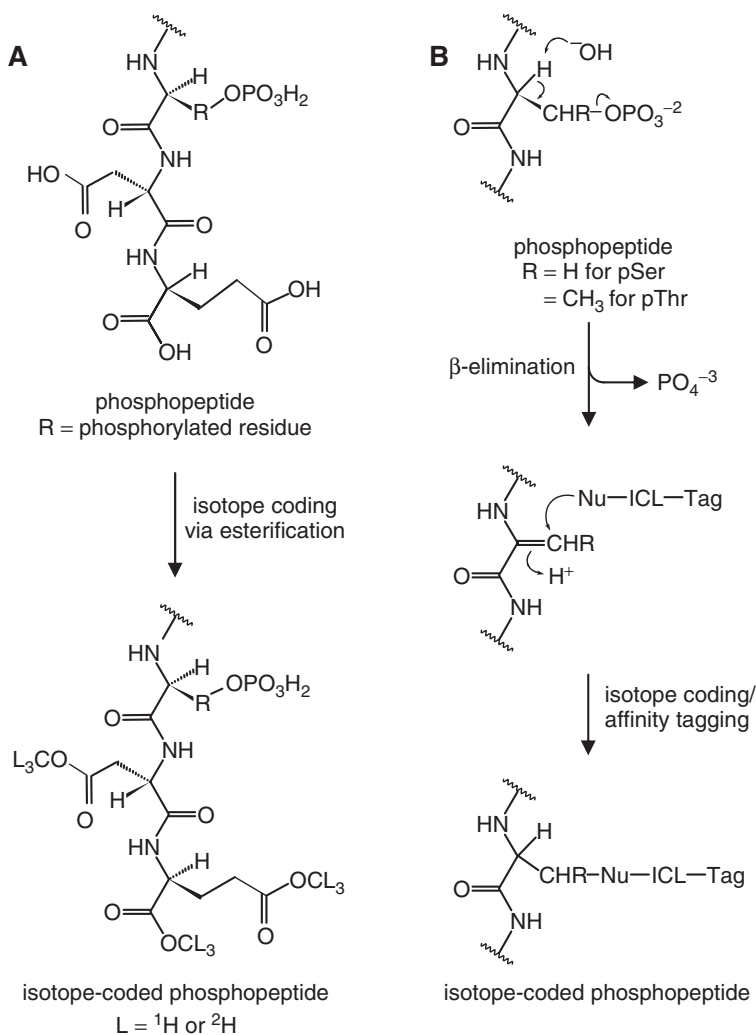


FIGURE 32.2. Stable isotope coding approaches for phosphoproteomics. **(A)** IMAC enrichment. The carboxylic acid groups of peptides (Asp, Glu, and C-terminus) are esterified in acidic methanol prior to IMAC using d_0 -methanol ($L = ^1\text{H}$) and d_3 -methanol ($L = ^2\text{H}$) to label control and experimental samples, respectively. **(B)** Stable isotope coding using β -elimination/Michael addition. Phosphoserine (pSer, $R = \text{H}$) and phosphothreonine (pThr, $R = \text{CH}_3$) peptides are subjected to β -elimination to remove the phosphate moieties and create electrophilic α , β -unsaturated double bonds that can be selectively labeled by reagents that contain a reactive nucleophile (Nu), an isotope-coded linker (ICL) for quantifying relative changes in protein phosphorylation, and an affinity tag or reactive functionality (Tag) to enable selective enrichment of labeled peptides. **(C)** LC-MS/MS analysis. Labeled peptides are identified by matching the product ion spectra (lower right) against a translated genomic database. By integrating the signal intensity for each isotope-coded peptide (lower left), the ratio of phosphorylation site abundances (L/H) between the two samples can be determined.

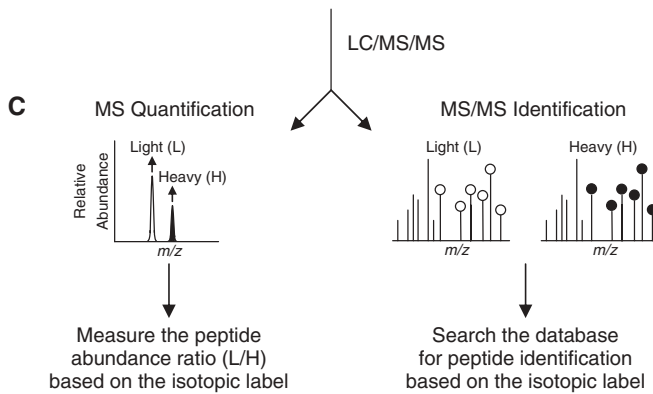


FIGURE 32.2. (Continued)

Goshe et al. [31, 32] developed a novel strategy to identify and quantify the extent of protein phosphorylation using a phosphoprotein isotope-coded affinity tag (PhIAT). In this method (Figure 32.2B, C), Cys-residues are blocked and the phosphate groups of pSer and pThr residues are removed by hydroxide ion-mediated β -elimination to produce thiolate-reactive sites. When comparing the relative phosphorylation states of phosphopeptides from two distinct samples, these thiolate-reactive sites are modified with isotopic versions of 1,2-ethanedithiol (EDT) that contain either four alkyl hydrogens (d_0 -EDT) or four alkyl deuteriums (d_4 -EDT). Once isotopically labeled, the d_0/d_4 -EDT-labeled residues are biotinylated using (+)-biotinyl-iodoacetamidyl-3,6-dioxaoctanediamine. The phosphorylated peptides are purified and concentrated using affinity chromatography and quantified by LC–MS/MS analysis. The relative stability of the PhIAT-label during CID enables the specific location of the site of phosphorylation to be identified from the product ion spectrum and permits the state of phosphorylation to be quantified [32]. An improved solid-phase-based version of PhIAT, termed phosphoprotein isotope-coded solid-phase tag (PhIST) [33], was developed in which isotope-coded solid-phase reagents containing either $^{12}\text{C}_6/^{14}\text{N}$ or $^{13}\text{C}_6/^{15}\text{N}$ and a UV-sensitive photolinker were used to increase efficiency and sensitivity for isolating and analyzing phosphopeptides. Because of its smaller tag and increased stability during CID, the PhIST labeling approach is more effective at identifying multiple phosphorylation sites on the same peptide while permitting the relative quantification of phosphorylation than PhIAT. Depending on the reaction conditions, the labeling of some protein samples can be substoichiometric, and difficulties associated with the low solubility of the EDT compound in aqueous solutions can promote protein precipitation. Recently, Goshe and colleagues [34] have addressed these issues by improving their PhIST approach to yield effective quantitative blocking of cysteinyl residues while promoting quantitative labeling of both pSer and pThr residues at the peptide level using Ba^{2+} -catalyzed β -elimination and Michael addition of (*R,R*)-DTT as the thiolate linker, circumventing the difficulties encountered with labeling at the protein level. Current work is underway to optimize these labeling strategies for identification and quantification of LRR RLK phosphorylation sites *in planta*.

Identification of Specific Phosphorylation Sites by LC-MS/MS

A phosphoproteomics approach to LRR RLK phosphorylation site analysis may reveal interesting trends and help identify individual LRR RLKs for further study, but is unlikely to yield a comprehensive picture of the phosphorylation status of any specific LRR RLK due to sample complexity. Critical residues may be transiently phosphorylated with varying stoichiometries, which makes identification of low-abundance phosphorylation sites difficult. In view of this, we are examining in detail the phosphorylation status of individual LRR RLKs in transgenic plants expressing FLAG-tagged proteins. Plants will be grown under conditions where that LRR RLK is expected to be phosphorylated, based on our preliminary proteomics approach.

Our model for this type of analysis is again the BRI1 receptor kinase. We chose the FLAG epitope tag because of its small size, lack of Ser and Thr residues, and low nonspecific binding of the commercial anti-FLAG antibody to plant proteins. A high degree of sample enrichment and purification can be obtained by immunoprecipitation of the tagged protein from solubilized microsomal membranes isolated from whole plants. After SDS-PAGE, the gel band containing the BRI1-FLAG protein is excised, digested with trypsin and analyzed by LC-MS/MS. Multiple approaches were required to achieve a high level of sequence coverage for BRI1-FLAG, as has been observed in many other phosphorylation studies [35]. Both Q-TOF and ion trap mass spectrometers were used, analyzing each sample with and without IMAC treatment. Some phosphorylated peptides were identified in all of our LC-MS/MS analyses, but others depended on the specific approach, suggesting that a single experimental method is unlikely to reveal all phosphorylation sites in a given protein due to varying sequence context of each tryptic peptide. Our experimental approach identified at least 11 sites of *in vivo* phosphorylation in BRI1-FLAG [21], and this approach is currently being applied to several other LRR RLKs including BAK1.

Functional Analysis of LRR RLKs by Mutant Complementation

Once the *in vivo* phosphorylation sites for a given LRR RLK are determined by LC-MS/MS analysis, *in vitro* mutagenesis of the original RLK-FLAG expression vector can be used to create a series of constructs in which each phosphorylated Ser or Thr residue in the specific LRR RLK is changed to an Ala residue. The ability of these constructs to rescue knockout or knockdown mutants of the particular LRR RLK is then compared to the corresponding wild-type construct. If a particular phosphorylated residue is critical for regulation of the signaling pathway, a Ser or Thr to Ala substitution at that site should result in a construct that cannot rescue the mutant phenotype, since an Ala residue will not be phosphorylated. This strategy has been successfully used to determine the functional importance of a number of *in vivo* phosphorylation sites of BRI1 (see below) and is applicable, in principle, to any LRR RLK where a mutant with an observable phenotype is available, provided that the mutant can be successfully transformed. While many T-DNA insertion lines for most LRR RLKs are publicly available, many have no phenotype, presumably because of functional redundancy. The NSF 2010 project directed by Dr. Frans Tax, University of Arizona (<http://www.mcb.arizona.edu/tax/2010/index.htm>), is identifying LRR RLKs

in which single mutants have strong phenotypes and is also examining double and triple mutants within LRR RLK subfamilies when single mutants have no phenotype. Such a genetic resource will be valuable for future functional analysis of *in vivo* phosphorylation sites of additional LRR RLKs.

Biochemical Characterization of Kinase Function

The cytoplasmic kinase domains of numerous LRR RLKs have been expressed as recombinant proteins in *E. coli*, and most often they behave as functional kinases *in vitro*. A number of important biochemical parameters can be assessed using recombinant kinase domains, and we have performed such analyses on BRI1 and BAK1 [5, 7]. Typically, the entire cytoplasmic domain, which consists of a juxtamembrane region, a catalytic kinase domain, and a short C-terminal domain, is expressed. These are generally produced as a fusion protein with a small N-terminal tag (e.g., FLAG- or His-tag) or an intact protein such as maltose-binding protein or GST. Our experience has been that FLAG-tagged versions of the cytoplasmic domains of the LRR RLKs generally express better than His₇-tagged counterparts and that GST-fusion proteins should be avoided because of possible dimerization of the GST protein *in vitro*.

There are two major functions of the recombinant LRR RLK cytoplasmic domains that can be assessed *in vitro*: autophosphorylation, which is generally studied by following the incorporation of radiolabel from [³²P]ATP into the recombinant protein, and trans-phosphorylation of either protein [37] or synthetic peptide [7] substrates. An advantage of the ³²P-labeling during autophosphorylation is that phosphoamino acid analysis can then be performed to ascertain whether Ser, Thr, or Tyr residues are involved. A drawback is that some recombinant protein kinases are highly autophosphorylated as isolated from *E. coli* [21, 36], and thus the “additional” phosphorylation that occurs *in vitro* may significantly underestimate the process. For kinase domains that will phosphorylate a substrate protein or synthetic peptide, various parameters of the reaction can then be characterized such as dependence on pH and the concentrations of substrate, ATP, and divalent cations. The availability of bacterial expression constructs for the entire LRR RLK family will allow us to assay these parameters on a comparative and family-wide basis.

32.3 EXPERIMENTAL RESULTS AND APPLICATIONS

Generation of a Gateway Compatible Construct Kit for LRR RLK Analysis

Our first objective was to clone 223 LRR RLK cytoplasmic domains (juxtamembrane region, catalytic kinase domain, and carboxy-terminal region) into three different Gateway-compatible vectors for *in vitro* biochemical analysis. The vectors selected were (a) pDONR/Zeo (Invitrogen, Carlsbad, CA) to create a family of entry clones that can be moved into any Gateway expression vector, along with two expression vectors that we constructed specifically for this project, (b) pFLAG-Gateway to generate a complete set of N-terminal FLAG-tagged proteins, and (c) pET15b-Gateway to create

a set of N-terminal His₇-tagged constructs. So far, we have successfully PCR amplified 217 LRR RLKs, and 197 of these have been cloned into pDONR/Zeo. Sequence analysis of the cytoplasmic domain in the vector construct shows that 175 of these exactly match the predicted amino acid sequence on the TAIR website (www.arabidopsis.org), with the remaining 12 being alternatively spliced or showing other polymorphisms. Of the 197 entry clones, 171 have been moved into the FLAG-tagged expression vector and 185 have been transferred to the His₇-tagged vector, generating a substantial resource for biochemical analysis.

For *in planta* functional analyses, the full-length cDNA sequences of the 223 LRR RLKs were amplified by RT-PCR and cloned into pDONR/Zeo. So far, nearly 90% of these have been cloned into the entry vector, and sequence analysis shows that 125 of these exactly match the amino acid sequence predicted by TAIR. Of these, 101 have been moved to a plant transformation vector that incorporates a C-terminal FLAG epitope tag on the LRR RLK. The *in vivo* interaction of BRI1 and BAK1 was demonstrated by co-immunoprecipitation in double transgenic plants expressing both BRI1-FLAG and BAK1-GFP [5]. To extend this type of interaction study to other LRR RLK pairs, two sets of Gateway compatible transformation constructs containing either FLAG-tagged or GFP-tagged LRR RLKs have been produced. The vectors developed specifically for this purpose contain either BASTA or kanamycin selectable marker genes to ensure future success of generating double transgenic plants harboring different LRR RLKs with different tags. All of the Gateway vectors generated for *in vitro* and *in vivo* LRR RLK analysis will be distributed to the scientific community via the *Arabidopsis* biological resource center (<http://www.biosci.ohiostate.edu/~plantbio/Facilities/abrc/abrhome.htm>).

Biochemical Properties of the BRI Kinase Domain

An initial biochemical characterization of the FLAG-BRI1 fusion protein found that BRI1 was autophosphorylated *in vitro* on at least 12 Ser and Thr residues [7]. FLAG-BRI1 also readily phosphorylates certain synthetic peptides *in vitro*. The best peptide substrate identified to date is the “BR13” peptide, GRJ**KK**IASVEJ**JK**K (J is nor-leucine, a stable Met analog; the phosphorylated Ser is underlined, and positive recognition elements are in bold italics; see below). This peptide is phosphorylated with an apparent K_m of ~82 μ M [7]. Variants of the BR13 peptide were prepared and tested as substrates, which yielded a preliminary phosphorylation motif of Basic-Basic-X-X-[ST]-X(3)- ϕ -Basic (where ϕ is a hydrophobic residue) (Figure 32.3A). Specificity information is also beginning to emerge in relation to autophosphorylation, because 16 sites have now been unambiguously identified by MALDI-TOF and LC-MS/MS analysis. Examination of the 10-residue window surrounding the autophosphorylated amino acid reveals some interesting trends. First, it is clear that certain amino acids have altered abundance relative to their total content within the cytoplasmic domain. For example, the entire BRI1 cytoplasmic domain contains 51 acidic and 52 basic residues; but surrounding the autophosphorylation sites, the basic residues are almost twice as abundant as acidic residues (25 versus 16, respectively). The distribution of basic and hydrophobic residues surrounding the autophosphorylated

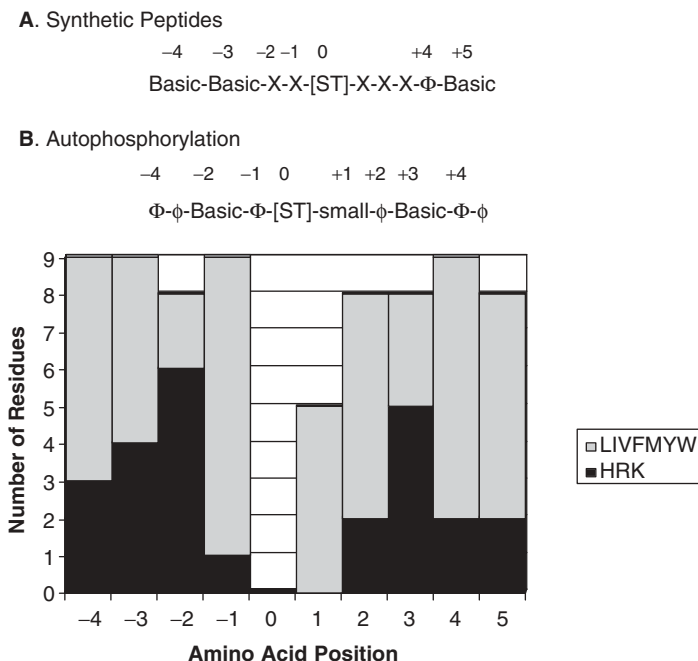


FIGURE 32.3. Preliminary BRI1-KD phosphorylation motifs. **(A)** Minimal motif based on synthetic peptide phosphorylation [7]. **(B)** Motif based on alignment of 16 unambiguous *in vitro* [7] (Hardin and Huber, unpublished data) and *in vivo* [21] autophosphorylation sites. The histogram shows the frequency of basic and hydrophobic residues from pS -4 to pS +5.

Ser or Thr is summarized in Figure 32.3B. In general, there was an inverse relationship between these two classes of amino acids at a given position. These results yield the preliminary autophosphorylation motif (Figure 32.3B) with relatively stronger preference for basic residues at the pS -2 and pS +3 positions, and hydrophobic residues at pS -1 and pS +4 positions. Thus, the motifs that emerge for autophosphorylation and *trans*-phosphorylation of synthetic peptide substrates are somewhat distinct but generally similar, with positioning of basic and hydrophobic residues apparently playing an important role.

Another interesting point is that the large polar residues, Asn and Gln, are relatively enriched within the 10-residue window surrounding autophosphorylation sites. The entire cytoplasmic domain of BRI1 contains about 3.7% Asn (which is rather typical for eukaryotic proteins), but the autophosphorylation “windows” contain 8.3% Asn. The residues are particularly enriched between pS -2 and pS +2, that is, directly surrounding the phosphorylated residue. The possible role of polar amino acids, especially Asn, in targeting is an exciting aspect that awaits future study. That additional residues in addition to hydrophobic and basic residues may play some role is suggested by the observation that the “BR12” synthetic peptide (AR**JRR**AASAAA**JK**A), which has all of the essential positive recognition elements (in bold italics) but with Ala

residues at the apparently nonessential positions, is a very poor substrate for recombinant BRI1-KD [7]. Thus, more work needs to be done to elucidate the recognition elements involved in BRI1 phosphorylation, and it is possible that large polar residues may unexpectedly play an important role.

Identification of *In Vivo* Phosphorylation Sites of BRI1 by LC–MS/MS

We used a variety of LC–MS/MS procedures to examine *in vivo* phosphorylation sites for BRI1-FLAG immunoprecipitated from *Arabidopsis* plants treated with 100 nM BL. Sequence coverage of the BRI1 cytoplasmic domain was over 91%, and six sites—S838, S858, T872, and T880 in the juxtamembrane region, T982 in the kinase domain, and S1168 in C-terminal region—were unambiguously identified as phosphorylated *in planta* [21]. Manual inspection of extracted ions chromatograms from the MS data of several ion trap LC–MS/MS analyses strongly suggested that there were two additional *in vivo* phosphorylation sites in the juxtamembrane region and that the peptide 1033-LMSAMDTLHSVSTLAGTPGYVPPEYYQSFR-1062, containing all of the six Thr and Ser residues within the BRI1 kinase subdomain VII/VIII activation loop, can be phosphorylated on up to three sites *in vivo*, consistent with our previous findings of autophosphorylation of three residues of the activation loop of the BRI1 kinase domain *in vitro* [7]. Thus, we have compelling evidence for at least 11 sites of *in vivo* phosphorylation in BRI1.

Phosphorylation of one to three residues in the kinase domain activation loop is often required for ligand-induced kinase activation, particularly in so-called RD type kinases that have an Arg immediately upstream of the invariant catalytic Asp in subdomain VIb [38]. Interestingly, alignment of 121 *Arabidopsis* RD-type LRR RLKs shows that over 99% have an S or T at the position aligning with BRI1 T1049 and more than 85% have an S or T corresponding to BRI1 S1044. Thus, these residues are likely candidates for BRI1 phosphorylation, and substitution of Ala for these residues by *in vitro* mutagenesis in BRI1 recombinant kinase domains causes loss of biochemical function *in vitro* [21]. While activation loop phosphorylation may be a general activation mechanism of the kinase domain, LRR RLKs function in many diverse physiological processes and mechanisms for transducing signals to pathway-specific cytoplasmic components are also required. Autophosphorylation of juxtamembrane and C-terminal regions, which show much less sequence conservation among receptor kinases than the activation loop, may create docking sites for specific LRR RLK substrates and thus be one way to achieve this selectivity [39]. Consistent with this idea, substitution of Ala for Ser or Thr in BRI1 juxtamembrane or C-terminal phosphorylation sites has little effect on general kinase autophosphorylation but significantly reduces peptide substrate phosphorylation *in vitro* [21].

Functional Characterization of Specific BRI1 Phosphorylation Sites *In vivo*

The BR-insensitive *bri1-5* mutant is a weak allele that shows intermediate dwarfism and a smaller rosette size than wild-type due to a C69Y amino acid substitution in the

extracellular domain of BRI1 [5]. Transgenic constructs containing the BRI1 promoter, the entire coding region and an in-frame C-terminal FLAG epitope tag fully rescue the *brl-5* phenotype to wild-type. Substitutions of Ala for specific Thr or Ser residues within the wild-type construct that were shown to be phosphorylated by LC–MS/MS analysis were generated and transformed into *brl-5*. Multiple independent transgenic lines for each construct were evaluated visually for phenotype compared with untransformed wild-type and the *brl-5* mutant and transgene expression and native BRI1 RNA levels were monitored. Substitutions in the activation loop dramatically affected the ability of the construct to rescue *brl-5*. T1039A and S1042A both showed a phenotype that was intermediate between *brl-5* and wild-type, while T1049A and S1044A were incapable of *brl-5* rescue in any of the transgenic lines examined, which is consistent with their nearly complete loss of kinase activity using *in vitro* biochemical assays [21].

Substitution of Ala for Thr or Ser in the juxtamembrane and C-terminal region of BRI1 did not appear to affect the general ability of the transgene to rescue the *brl-5* mutant to an apparently normal wild-type size, consistent with our *in vitro* observation that none of these mutations affected general autophosphorylation of the BRI1 cytoplasmic domain. However, based on our *in vitro* peptide substrate assays, we expected that some of these mutations might show an altered phenotype due to inability to phosphorylate a downstream substrate to the extent required for normal BR signaling. It is possible that the juxtamembrane and/or C-terminal mutations may have subtle phenotypes due to altered phosphorylation of substrates in a specific branch pathway of BR signal transduction, and we are testing this possibility using a detailed morphometric analysis of several growth parameters in homozygous T₃ lines for each mutation. Based on *in vitro* and *in vivo* data and on conservation of function in other LRR RLKs, we propose that BL-dependent phosphorylation of T1049 and S1044 are required for activation and function of BRI1 *in planta*, while phosphorylation of individual juxtamembrane and carboxy-terminal Ser and Thr residues might affect further phosphorylation of BRI1 cytoplasmic substrates, but apparently is not required for general kinase activation [21].

32.4 CONCLUSIONS AND FIVE-YEAR VIEWPOINT

Recent advances in proteomics, MS analysis, and high-throughput cloning have facilitated a global comparative analysis of biochemical function of the entire LRR RLK family in *Arabidopsis*. The genetic resources and completed genomic sequence of this model plant, coupled with increased sensitivity of LC–MS/MS approaches and reliable quantitative methods for following phosphorylation status at specific residues *in vivo*, will allow building of a comprehensive database of LRR RLK phosphorylation sites as well as analyzing the biological significance of these phosphorylation events. Over the next five years we plan to significantly increase the number of LRR RLKs included in this database, which should enhance our understanding of the molecular mechanisms involved in the control of plant growth and development by LRR RLKs. This family-wide analysis of the biochemical properties of LRR RLK kinase domains

may also show whether differential kinase activity, autophosphorylation sites, and substrate preferences will allow a grouping of kinases that correlates with the phylogenetic arrangement of subfamily categories determined by kinase domain sequence alignment. The LRR motif is thought to be involved in PPIs, and it would not be surprising if many LRR RLKs formed a variety of heterodimers among family members, which might be essential for signaling responses to different ligands. Thus, a tool set of LRR RLK clones with different epitope tags will be useful for interaction studies both *in vitro* and *in vivo*. Characterization of the BRI1/BAK1 model system has shown that plant receptor kinases share the general paradigm of animal receptor kinase function, including ligand-dependent heterodimerization and phosphorylation on specific residues of the cytoplasmic kinase domain. We plan to expand the BRI1/BAK1 approach to other LRR RLKs of both known and unknown function to increase our understanding of this large family of critical signaling proteins.

ACKNOWLEDGMENTS

The authors would like to thank the research agencies of North Carolina State University and the North Carolina Agricultural Research Service for continued support of biological mass spectrometry research. Portions of this work were supported by grants from the NSF (MCB-0419819) and the United States Department of Agriculture (NRI 2004-35304-14930).

REFERENCES

1. Shiu SH, et al. 2004. *Plant Cell* **16**:1220–1234.
2. Shiu SH and Bleeker AB. 2001. *Proc Natl Acad Sci USA* **98**:10763–10768.
3. Torii KU. 2004. *Int Rev Cytol* **234**:1–46.
4. Li J, Chory J. 1997. *Cell* **90**:929–938.
5. Li J, et al. 2002. *Cell* **110**:213–222.
6. Nam KH and Li J. 2002. *Cell* **110**:203–212.
7. Oh MH, et al. 2000. *Plant Physiol* **124**:751–766.
8. Wang ZY, et al. 2001. *Nature* **410**:380–383.
9. Johnson KL and Ingram GC. 2005. *Curr Opin Plant Biol* **8**:648–656.
10. Marmagne A, et al. 2004. *Mol Cell Proteomics* **3**:675–691.
11. Nuhse TS, et al. 2004. *Plant Cell* **16**:2394–2405.
12. Alexandersson E, et al. 2004. *Plant Cell Physiol* **45**:1543–1556.
13. Nelson CJ, et al. 2006. *Mol Cell Proteomics* **5**:1382–1395.
14. Einhauer A and Jungbauer A. 2001. *Biochem Biophys Methods* **49**:455–465.
15. Curtis MD and Grossniklaus U. 2003. *Plant Physiol* **133**:462–469.
16. Clouse SD and Sasse JM. 1998. *Annu Rev Plant Physiol Plant Mol Biol* **49**:427–451.
17. Kinoshita T, et al. 2005. *Nature* **433**:167–171.
18. Bishop GJ. 2003. *J Plant Growth Regul* **22**:325–335.

19. Vert G, et al. 2005. *Annu Rev Cell Develop Biol* **21**:177–201.
20. Russinova E, et al. 2004. *Plant Cell* **16**:3216–3229.
21. Wang X, et al. 2005. *Plant Cell* **17**:1685–1703.
22. Luche S, Santoni V and Rabilloud T. 2003. *Proteomics* **3**:249–253.
23. Blonder J, et al. 2004. *Proteomics* **4**:31–45.
24. Larsson C, Sommarin M, and Widell S. 1994. *Methods Enzymol* **228**:451–469.
25. Mitra, SK, et al. 2007. *J Proteome Res* **6**:1933–1950.
26. Ficarro SB, et al. 2002. *Nat Biotechnol* **20**:301–305.
27. Ficarro S, et al. 2003. *J Biol Chem* **278**:11579–11589.
28. Ndassa YM, et al. 2006. *J Proteome Res* **5**:2789–2799.
29. Gerber SA, et al. 2003. *Proc Natl Acad Sci USA* **100**:6940–6945.
30. Ross PL, et al 2004. *Mol Cell Proteomics* **3**:1154–1169.
31. Goshe MB, et al. 2001. *Anal Chem* **73**:2578–2586.
32. Goshe MB, et al. 2002. *Anal Chem* **74**:607–616.
33. Qian WJ, et al. 2003. *Anal Chem* **75**:5441–5450.
34. Soderblom EJ, et al. 2005. *Int J Neuroprotec Neuroregen* **1**:91–97.
35. Loyet KM, Stults JT, and Arnott D. 2005. *Mol Cell Proteomics*: R400011–MCP400200.
36. Hardin SC, et al. 2006. *Plant Biol SUPPL*:53–54. <http://abstracts.aspb.org/pb2006/public/P36/P36018.html>.
37. Ehsan H, et al. 2005. *Plant J* **43**:251–261.
38. Johnson LN, Noble ME, and Owen DJ. 1996. *Cell* **85**:149–158.
39. Pawson T. 2004. *Cell* **116**:191–203.

TIME TO SEARCH FOR PROTEIN KINASE SUBSTRATES

Birgit Kersten

33.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Reverse protein phosphorylation is a highly conserved mechanism for post-translational regulation of protein function, which has been found in all prokaryotes and eukaryotes examined (see also Chapter 29). Phosphorylation is mediated by protein kinases, which transfer the γ -phosphate of ATP on amino acid residues of proteins, and can be reversed by protein phosphatases. The phosphorylation state of a protein has often profound effects on its activity, stability, structure, and interaction with other proteins or biomolecules [1]. Virtually all cellular processes are regulated in one or multiple ways by phosphorylation and dephosphorylation. Phosphorylation plays a central role, especially in cell signaling in diverse organisms including plants [2]. The phosphorylation of proteins in plants has been found to be connected with the reaction of the organism toward different internal and external factors, such as light, invasion of pathogens, hormones, temperature stress, and shortage of nutrition [3].

Current annotation of protein families in *Arabidopsis* predicts approximately 1053 kinases, representing about 4% of all *Arabidopsis* proteins (Pfam; http://www.sanger.ac.uk/cgibin/Pfam/genome_view.pl?ncbi=3702) [4]. In contrast, the human genome

with approximately the same number of genes as that of *Arabidopsis* is estimated to encode for approximately 500 protein kinases, accounting for only almost 2.0% of its genome [5, 6]. Why do plants need a comparable high number of protein kinases? Analysis of these kinases, knowing which kinases are activated during a biological response and identifying their specific substrates, is important for a better understanding of the regulation of a wide range of biological processes in plants and to increase our knowledge of what makes plants different from other organisms.

Despite recent advances that have been made in phosphoproteomics [7] (see Chapter 29), we know very little about kinase activity pattern, substrates, and the elements mediating specificity in plants, compared to the status in mammals. Therefore, the assignment of phosphoproteins to specific kinases and the search for novel substrates of plant kinases are great challenges.

Several criteria for establishing that a protein is a substrate of a given kinase were defined by Berwick and Tavare [8] comprising the ability of the kinase to phosphorylate the substrate *in vitro* and the dependence of substrate phosphorylation on kinase activity *in vivo*. In a recent paper, Peck [4] described very detailed strategies for monitoring kinase activity, investigating kinase-substrate specificity, examining phosphorylation *in planta* and determining phosphorylation sites in a protein. It is underlined that it is very important to establish, by different assays and strategies, links in the following chain of a signal transduction: (i) environmental or internal triggers (e.g., specific abiotic or biotic stresses), (ii) activation of a kinase, (iii) specific phosphorylation of a substrate by the kinase, and (iv) specific cellular response. To demonstrate that a kinase is involved in a signal transduction, it is essential to demonstrate that the activity of the kinase increases during the biological response. One way is to immunoprecipitate the kinase from plant cell extracts, preferably using a specific antibody against the native kinase. Then the activity of the immunoprecipitated kinase is assayed using a reporter protein—for example, a “generic substrate” such as MBP. Once a kinase has been determined to be activated during a response, the hunt begins for substrates of this kinase [4]. Because the identification of kinase substrates *in vivo* continues to be laborious and time-consuming, it is more efficient to identify substrate candidates in an efficient manner *in vitro* before proceeding to the *in vivo* step. *In vitro*-identified substrate candidates and their phosphorylation sites have to be confirmed *in vivo*. Different *in vitro* and *in vivo* methods and strategies used in search for kinase substrates, their increasing application in the plant field, and the obtained results will be reviewed in this chapter.

33.2 METHODOLOGY AND STRATEGY

***In Vitro*-Strategies and Methods for Identification of Candidate Substrates**

Overview. Classical *in vitro* methods for the identification of substrate candidates for protein kinases are, for example, the identification of protein interaction partners of a given kinase. These candidates are then tested for phosphorylation by the kinase *in vitro*. For this purpose the purified candidate protein is often incubated with

a protein kinase in solution, usually in the presence of radioactive ATP and MgCl_2 followed by SDS-PAGE of the proteins and the detection of the phosphorylated substrate by autoradiography [9].

In-gel kinase assays are often used to identify and analyze a specific kinase for a given substrate in crude extracts compared to other kinases in the extract. In this assay, the putative substrate is casted in the SDS gel. After applying the kinase, for example, with the cellular extract onto the gel and protein separation, the proteins, including kinases, are refolded in the gel prior to an in gel-kinase assay with radioactive ATP. Specific phosphorylation of the substrate results in a radioactive band at a position that corresponds to the MW of the phosphorylating kinase. This approach was used, for example, to demonstrate that the *Arabidopsis* MAP kinase, AtMPK6 but not AtMPK3, specifically phosphorylates 1-aminocyclopropane-1-carboxylic acid synthase 6 (ACS6) *in vitro* [10]. One limitation of using in-gel kinase assays is that not all kinases refold into an active form after separation by SDS-PAGE.

Although all these techniques are extremely useful, high-throughput phosphorylation methods are in great demand to discover novel substrates. Over recent years, several proteomics approaches have been established for phosphorylation screening in a high-throughput manner and find increasing application in the plant field. *In Vitro* kinase assays may be performed in solutions of crude cell extracts (see *Kinase Assays in Cellular Extracts*). Furthermore, array-based methods are more and more applied for *in vitro* phosphorylation screening in an unbiased fashion by screening huge numbers of proteins for phosphorylation in parallel (see *Solid-Phase Phosphorylation Screening*).

In Vitro Protein-Protein Interactions Screening Methods. High-throughput methods to screen for PPI may be applied to find novel candidate substrates. The development of the Y2H system [11] and its adaptation to high throughput [12] are important steps toward the efficient identification of such protein interactions in living yeast [13, 14]. In a typical Y2H search, a hybrid protein consisting of the fusion of a DNA-binding domain (DBD) of a TF and a protein of interest (bait) is assayed against a single protein (prey) or a prey-library. Prey proteins are expressed as fusions with a transcriptional activation domain (AD, commonly derived from the yeast protein GAL4). Via the interaction of bait and prey protein, an artificial TF (DBD+AD) is reconstructed triggering the activation of a reporter system in a suitable yeast strain. Some novel plant kinase substrates could be identified starting from a Y2H-PPI, such as the *Arabidopsis* MAPK4 substrate MKS1 [15] or the tobacco TF WRKY1, which has been revealed as a potential substrate of SIPK, a MAPK of tobacco, by a targeted Y2H approach [16].

Solid-phase screening of non-arrayed expression cDNA libraries in phage vectors by far-Western is another powerful method for the identification of PPIs. By interaction screening of such a library from *N. tabacum* with a radioactive labeled TMV-MP, a clone encoding a homologue of the unique protein kinase, RIO (right ORF), has been identified. The kinase was shown to phosphorylate the TMV-MP [9].

In recent years, different array-based methods are increasingly used for the global analysis of PPIs, which can be applied also for the identification of kinase interactions

with substrate proteins. Protein microarrays are a powerful tool in this respect [17]. To manufacture protein microarrays, purified proteins are spotted by contact or piezo microarrayer onto coated glass slides. After incubation with the fluorescently labeled kinase, washing, and scanning, interaction partners can be identified. For interaction studies, 5600 purified yeast proteins were arrayed and screened for their ability to interact with other proteins [17]. Using this approach, many unknown CaM-interacting proteins were identified. One major drawback, however, is the necessity of using purified proteins in most of the microarray-based methods. Protein purification is usually difficult, time-consuming, and expensive. Besides different microarray-based *in vitro* transcription/translation strategies [18, 19], recently the design and application of a membrane-based array technology has been reported in a human study [20]. This technology allows the identification of PPIs with crude bacterial cell extracts without purification of the recombinant proteins and is also feasible for *E. coli* clones expressing plant proteins.

Applying PPI screening methods, one has to take into account that kinase–substrate interactions are often not detectable by PPI assays. The results of a global study on protein phosphorylation in yeast support this expectation [21]. In this study, only a low percentage of the substrate-kinase pairs identified by phosphorylation on protein microarrays were also present in the PPI network drawn from the literature. Nevertheless, several new plant kinase substrates could be identified starting from an *in vitro* PPI as mentioned [15, 16].

Kinase Assays in Cellular Extracts. *In Vitro* phosphorylation reactions can be performed in native whole-cell extracts, thus presenting an environment that closely resembles the *in vivo* state during the kinase reaction. Reactions can be run either by endogenous kinases [22] or by a purified protein kinase that is added to the reaction mixtures [23] in the presence of MgCl_2 and $[\gamma\text{-}^{32}\text{P}]\text{ATP}$. The resultant phosphorylated proteins are detected after SDS-PAGE by autoradiography. The assignment of a specific kinase to the *in vitro*-identified substrates is often not possible by running the kinase reactions only with the endogenous kinases. Even exogenous application of the kinase to the reaction mixture makes the assignment complicated because of the background phosphorylation by endogenous kinases. Often the discovery of a specific substrate–kinase link is only possible using differential display–strategies by comparing the phosphorylation patterns of different treated samples. Asano et al. [23] performed *in vitro* phosphorylation by addition of recombinant rice SPK (serine/threonine protein kinase), a calcium-dependent kinase, to crude extracts of immature rice seeds in the presence and absence of calcium. In the reaction run in the presence of calcium, the researchers could detect enhanced phosphorylation of a protein band corresponding to sucrose synthase, identifying this protein as a potential substrate. Beside the background phosphorylation problem, the following limitations of *in vitro* kinase approaches in whole-cell extracts have to be mentioned: (i) the high ATPase activity in cell lysates, (ii) the high complexity of the sample, which renders subsequent substrate identification difficult, and (iii) the high dynamic range of protein concentration in the native extracts, thus hampering detection of the substrate phosphorylation for low concentrated proteins.

An ingenious “chemical genetic” approach aiming to overcome background phosphorylation problems in crude cell lysates and to attribute phosphorylation of a protein to a particular kinase has been pioneered in Kevan Shokat’s laboratory [24]. This method allows for displaying direct substrates of a modified kinase of interest in the lysates. Therefore, a functionally silent mutation is engineered into the ATP-binding site of the kinase of interest, creating an enlarged but functional ATP-binding pocket that uniquely enables the mutant kinase to use an ATP analogue with a bulky side chain. If the radiolabeled ATP analogue is then provided in the kinase reaction mixture, only the mutant kinase will create radiolabeled phosphorylation products. This technique has been used, for example, to find novel substrates of ERK2 [25].

Solid-Phase Phosphorylation Screening. Tony Hunter’s group developed a powerful solid-phase phosphorylation screening method utilizing λ -phage cDNA expression libraries on nitrocellulose membranes, which are incubated with purified protein kinase and [γ - ^{32}P]ATP [26, 27]. Phages that encode potential substrates of the kinases are detected by autoradiography. The identity of the substrates is then determined by sequencing the cDNA insert. The Hunter method has been successfully applied, for example, to find MAPK-interacting kinase as a substrate of ERK1 and the p38 MAPK [26].

Protein microarrays containing recombinant proteins have become an important high-throughput tool in the proteomic field, because they allow fast and easy analysis of thousands of addressable immobilized proteins in parallel [28–32]. Due to their small size (size of a microscopic slide), only minute amounts of spotted proteins and analytes are needed [33]. Protein microarrays may be applied for different kinds of interaction studies such as PPIs or protein–DNA interactions [34] or to study the interaction of the immobilized proteins with small molecules [35]. The application of protein microarrays also expanded into the plant field, when the first *Arabidopsis* protein microarrays were generated to profile plant antibodies and sera with regard to their specificity and cross-reactivity [36].

Beside interaction analysis, different types of modification studies can be performed using protein microarrays, such as to study the phosphorylation of the immobilized proteins by kinases. The general capacity of protein microarrays to detect protein phosphorylation by kinases was demonstrated in the pioneering studies of MacBeath and Schreiber [37] and Zhu et al. [38]. MacBeath and Schreiber used microarrays coated with a BSA-*N*-hydroxysuccinidimide monolayer to analyze protein phosphorylation on these planar protein microarrays. They spotted the substrates of three well-known kinase–substrates pairs in quadruplicates onto three microarrays, and each slide was incubated with one of the three kinases in the presence of radioactive ATP [37]. Zhu and colleagues used polydimethylsiloxane-covered microarrays containing nanowells to analyze the activity and specificity of 119 yeast kinases for 17 different substrates. They immobilized one substrate per microarray. To analyze the substrate specificity of the kinases, they applied different kinases into the nanowells of a microarray together with radioactive ATP. By phosphorimaging to detect the radioactive signal, they identified 32 kinases, which preferentially phosphorylated one or two substrates [38].

Recently we performed the first protein microarray-based study in search for novel candidate substrates of protein kinases in *Arabidopsis* [39] after a proof-of-principle study in barley [40, 41]. With the barley study, we successfully established a protein microarray-based kinase assay. We applied the method to identify potential substrates of a barley casein kinase2 α (CK2 α) out of 768 recombinant barley proteins expressed from a cDNA expression library [40]. Besides known substrates of CK2 α , such as calreticulin, targets *hithero* not described in the literature were identified [40]. In the barley study, the selection of potential substrates was performed only based on qualitative criteria: A protein was regarded as potential substrate, in the case where all four of the corresponding quadruplicates in two independent experiments were detectable.

This assay has been further developed in the *Arabidopsis* study by increasing the density and therewith the amount of the arrayed proteins in order to design a high-throughput assay [39, 42]. Furthermore, steps for signal quantification, evaluation, and threshold-based selection of potential substrates were introduced to increase the reliability of the method. The advanced assay was applied addressing the challenging task of identifying potential substrates for *Arabidopsis* MPK3 and MPK6 [39]. As a source of recombinant *Arabidopsis* proteins, cDNA expression clones have been generated according to the workflow depicted in Figure 33.1. Briefly, an ordered cDNA library consisting of 38,000 clones was generated from *Arabidopsis* inflorescence meristem in an *E. coli* expression vector. Five thousand putative His-tagged expression clones were re-arrayed from this library after anti-His screening of protein filters that contained all library clones. Based on sequence analysis of the 5000 clones, an almost nonredundant clone set consisting of 1500 clones was created from the expression library [39] and enriched with ~ 200 full ORF expression clones [36] (1700 clones in total). Access to the clones including detailed clone descriptions has been recently provided via the GABI Primary Database (<http://www.gabipd.de>) at the RZPD German Resource Center for Genome Research Berlin. From all of the 1700 clones, proteins were then expressed and purified in 96-well format.

To generate protein microarrays (see Figure 33.2), all *Arabidopsis* proteins were immobilized alongside with different controls, such as BSA and sample buffer using contact printing onto FASTTM slides. Protein microarrays were screened with an anti-RGS-HIS₆ antibody to check the immobilization of the recombinant proteins on the microarrays (Figure 33.2). To identify potential substrates for a given kinase, two experiments with two microarrays each were performed for every kinase. For the kinase assays, protein microarrays were incubated with GST-tagged MPKs in the presence of radioactive ATP. Signals were detected by phosphorimaging. Figure 33.2 gives a typical result for a kinase tested. Signals were quantified and the signal intensities were evaluated in field-to-field and slide-to-slide correlations [39, 42]. Potential substrates were selected based on well-defined thresholds; for example, 48 substrate candidates were selected for MPK3 out of 1700 proteins. Nearly all of them were confirmed by an *in vitro* assay with the recombinant candidate proteins, which were refolded on NiNTA-beads prior to a kinase assay in solution [39].

Using a similar approach with radioactive detection, Ptacek and co-workers [21] recently assayed different yeast kinases on protein microarrays containing about 4400 recombinant yeast proteins. The phosphorylation assays mentioned above were

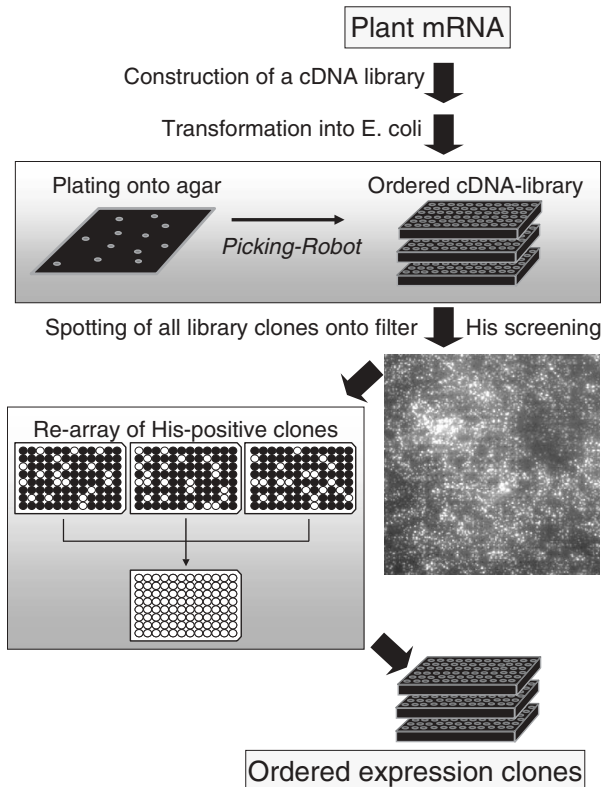


FIGURE 33.1. Workflow for the generation of ordered plant expression clones. Starting with plant mRNA, cDNA is generated and cloned into an *E. coli* expression vector for expression of His-tagged proteins. The cDNA library is transformed into *E. coli* and plated onto agar. The *E. coli* colonies are picked from the agar into the wells of 384-well plates. All clones of the ordered cDNA library are spotted onto PVDF filter. The filters are screened with an anti-His antibody to detect putative expression clones, expressing His-fusion proteins. All His-positive clones are re-arrayed into new 384-well plates. Ordered expression clones can be expressed and proteins can be purified via His-tag.

performed by using radioactive ATP. First attempts to detect phosphorylation status on microarrays with fluorescence were made by Martin et al. [43] using Pro-Q DPS. The main advantage of using a phosphostain would be that a common microarray scanner could be used for the quantitative evaluation of the microarrays. Unfortunately, the radioactive detection is still more sensitive. Nevertheless, a number of promising studies have been performed recently, using Pro-Q DPS for the detection of phosphorylation events on microarrays [44].

As a whole, solid-phase phosphorylation screens are very informative and allow for short-listing candidate substrates of plant kinases. The following are some drawbacks of these methods: (i) Protein kinases could show promiscuous activity *in vitro*,

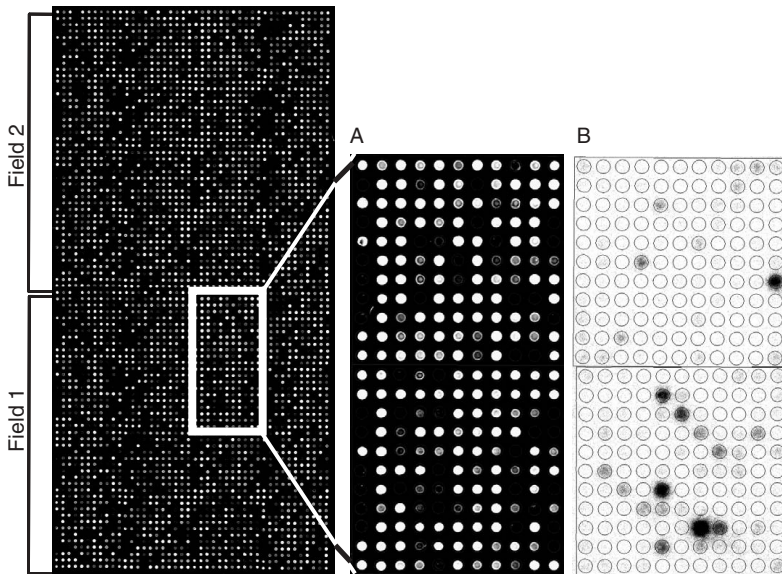


FIGURE 33.2. *Arabidopsis* protein microarray containing 1700 proteins. Purified proteins were spotted into two identical fields of FAST slides. Detection of recombinant proteins on a microarray with an anti-RGS His antibody (left) and section of the same array (A). (B) Phosphorimage of a microarray section after performing the microarray-based kinase assay. Images are taken from reference 39, copyrighted by *Molecular and Cellular Proteomics* 2005. *Arabidopsis* clones used in this study are accessible and described in detail via GABI Primary Database (<http://www.gabipd.de/>).

(ii) proteins expressed in prokaryotic systems might not fold appropriately, (iii) co-factors or binding proteins which are important for the *in vivo* phosphorylation may be missing under *in vitro* conditions, and (iv) the kinase could be localized in a different cellular compartment as the candidate substrate, being unable to phosphorylate it *in vivo* [1]. For all these reasons, follow-up *in vivo* experiments are essential to evaluate the physiological relevance of the potential substrates identified by solid-phase phosphorylation screens.

In Vitro Analysis of Phosphorylation Sites Using Peptide Arrays. Besides different MS methods, which can be applied to detect the sites of phosphorylation (see Chapter 29), peptide arrays with immobilized or synthesized peptide libraries are in use [45]. If the putative substrate is known, then knowledge-based libraries comprising overlapping peptides with sequences derived from this protein can be applied to identify the site(s) of phosphorylation (phosphorylated peptide/s) by incubation of the peptide arrays with the active kinase. In a plant study, the phosphorylation site for the calcium-dependent kinase from maize seedlings was identified and optimized using a knowledge-based library [46]. Therefore, peptides with substitutions in each position were derived from the phosphorylatable sequence of the known substrate sucrose

synthase. A consensus sequence was identified resembling the primary structure of the sucrose synthase phosphorylation site and showing similarities to the analogous sequence proposed for mammalian protein kinase C.

The knowledge on phosphorylation sites for specific kinases is important to design phosphopeptides for the generation of anti-phospho antibodies (see also *Methods to Analyze the Phosphorylation State of Proteins in Plant Extracts*), which detect specifically the phosphorylated protein. Furthermore, this knowledge can be applied in computational approaches for phosphosite prediction in protein sequences (see also Chapter 29).

***In Vivo* Approaches: Verification of *In Vitro* Results**

Overview on *In Vivo* Strategies. As described above, *in vitro* results can be very good indicators of specific phosphorylation of a protein by a given kinase. However, these data should be supported by experiments that analyze the phosphorylation status of the protein *in planta* [4]. It has to be demonstrated that the phosphorylation of the candidate protein depends on the activation of the respective kinase and shows a similar kinetics than the kinase activation in the scope of the analyzed biological response. Furthermore, *in vitro*-determined phosphorylation sites of the substrate candidates have to be confirmed *in planta*.

To study the activation of a kinase of interest and to reveal the downstream substrates in the scope of a specific biologic response, plants can be treated by a stress to induce the respective response. After demonstrating the activation of the respective kinase in the scope of the response, the phosphorylation state of the candidate substrates can be detected by different methods which will be mentioned below. Additional lines of evidence are, for example, (i) to demonstrate that the phosphorylation of the candidates will not occur in genotypes lacking the kinase activity (kinase knockout mutants), (ii) to detect phosphorylation of the candidates in plants overexpressing the kinase of interest, and (iii) to modulate kinase activity by site-directed mutagenesis of Ser and Thr in the candidate substrate [4, 10].

Methods to Analyze the Phosphorylation State of Proteins in Plant Extracts. Several methods for the enrichment and detection of phosphoproteins as well as the identification of the phosphorylation sites have been described in detail (see Chapter 29). The analysis of the phosphorylation state of specific candidate proteins is often hampered by the high complexity, even of samples enriched for phosphoproteins. If the substrate candidate is known and a specific antibody is available, then the protein may be enriched by immunoprecipitation prior to MS analysis for the detection of the phosphorylation site. Another possibility is to overexpress the candidate fused to an affinity tag in plants which allows subsequent immunoprecipitation of the candidate using a tag-specific antibody for further analysis. Specific antibodies against the candidate can be also applied to detect the respective protein and its modified forms in 2D gels of plant samples.

If the phosphorylation site is known, then it is possible to generate phospho-specific antibodies, which allows monitoring the phosphorylation state of the protein

in complex plant samples. One fashionable method in this respect is the application of protein microarrays. Two main types of protein microarrays are currently in use to study the phosphorylation of proteins in crude cell or tissue extracts: antibody microarrays (AMAs) and reverse protein microarrays (RPMAs) [31]. AMAs contain specific antibodies, antibody mimics, or other affinity reagents such as aptamers on their surface. To study protein phosphorylation, AMAs (e.g., with immobilized phospho-antibodies) that are incubated with the labeled crude extract can be used. The RPMAs are generated by spotting the complex protein samples themselves, which are then probed with phospho-specific antibodies to detect phosphorylated proteins in these samples. This type of microarray is very useful to monitor the phosphorylation state of a protein in different samples in parallel—for example, in the scope of kinetic studies. Chan and co-workers studied the kinetics of MAPK phosphorylation during the activation of Jurkat T cells using RPMAs. Lysates generated from the cells over a 30-min time period after activation were used to fabricate the RPMAs [47]. The arrays were incubated with Cy3-labeled anti phospho-MAPK antibodies to detect the level of MAPK phosphorylation. In a recent study, a combination of 2D liquid-phase protein separation with RPMA analysis of the fractions has been used for global profiling the phosphoproteome in human breast cancer cells [48]. Despite this effort, plant applications of these protein microarray-based methods have not been reported so far.

33.3 EXPERIMENTAL RESULTS ON PLANT MAPK DOWNSTREAM SIGNALING AND APPLICATIONS

Various types of kinases have been described from plants and classified into different groups [3]. One of the largest and most important families of Ser-Thr kinase is the MAPK family. MAPKs are the last component of the “three-kinase” modules of MAPK cascades [49]. This cascade consists of at least a MAPKKK, a MAPKK and finally a MAPK module (Figure 33.3). In the process of signal transduction the Ser/Thr protein kinase MAPKKK is phosphorylated by a receptor. The activated MAPKKK phosphorylates then the subsequent dual specific MAPKK, which in turn activates the MAPK by phosphorylation of a threonine as well as a Tyr residue in the “activation loop.” Once the terminal MAPK is activated, it can then phosphorylate downstream substrates [50].

MAPK cascades are universal and highly conserved signal transduction modules in eukaryotes, including yeasts, animals, and plants, and mediate the intracellular transmission and amplification of extracellular stimuli, resulting in the induction of appropriate biochemical and physiological cellular responses [51, 52]. Activation of MAPK cascades is an important mechanism for stress adaptation by control of gene expression. In plants, MAPK signaling has been implicated in abiotic as well as biotic stress situations and is associated with various physiological, developmental, and hormonal responses [53]. In certain plant species, this activation is an important component in host and non-host resistance against several pathogens and exhibits strong similarity to the innate immune protection systems of mammals and *Drosophila* [51].

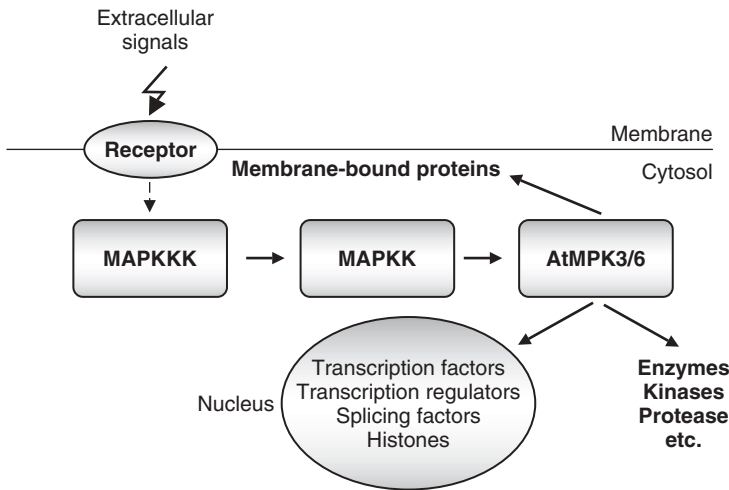


FIGURE 33.3. Hypotheses on *Arabidopsis* MPK3/6 downstream signaling. Potential substrates of these kinases which were identified by a protein microarray-based kinase assay [39] are depicted on a scheme of the MAPK signaling cascade.

In *Arabidopsis*, a complete MAPK cascade following perception of a bacterial flagellin, has been elucidated [49] (Figure 33.3). Downstream of the flagellin receptor, a LRR receptor kinase, the cascade consists of MEKK1 (MAPKKK), MKK4/MKK5 (MAPKKs), and MPK3/6 (MPKs); signaling via this cascade results in up-regulation of WRKY22/WRKY29 transcription factor gene expression. Despite this effort, nothing was known about the phosphorylation events and their influence on gene expression downstream of activated MAPKs. *In vitro* results obtained from protein microarray experiments [39] indicated that the MAPK signaling in plants is very complex and not restricted to the transcriptional level (see Figure 33.3) [50]. Interestingly, the TRA-like splicing factor (At1g07350), which has been recognized as an MPK3-substrate on the microarrays, has been identified to be phosphorylated by the group of Heribert Hirt in a recent *in vivo* study on *Arabidopsis* proteins involved in RNA metabolism [54]. Moreover, ACS-6, which has been detected as an *in vitro* substrate of MPK6 on the protein micorarrays [39], has also been identified as MPK6 substrate *in vivo* and therewith as the first plant MAPK substrate by Liu and Zhang [10]. This result demonstrates that the microarray-based kinase assay has the potential to identify substrates with physiological relevance. Liu and Zhang found that selected isoforms of ACS, which is the rate-limiting enzyme of ethylene biosynthesis, are MPK6 substrates [10]. Phosphorylation by MPK6 leads to the accumulation of ACS proteins and to elevated levels of ACS activity *in vivo*, consequently increasing ethylene production. Recently the AtMPK4 substrate, called MKS1 (MAP kinase substrate 1), was identified. This protein was shown to interact with the WRKY TFs, WRKY25, and WRKY33 [15]. As the authors could show in their study, MKS1 may contribute to MPK4-regulated defense activation by coupling the kinase to a specific WRKY TF.

33.4 CONCLUSIONS

Today, several advances have been made in phosphoproteomics technologies, including enrichment, detection, and quantification of phosphoproteins as well as mapping of their phosphorylation sites. Nevertheless, the assignment of the identified phosphoproteins to specific kinases and the search for novel substrates in an effective manner are very challenging. *In vitro* screening methods for protein phosphorylation performed on protein microarrays or in solution as well as PPI studies succeeded in the identification of novel substrate candidates for some plant kinases *in vitro*. Follow-up *in vivo* experiments are essential to evaluate the physiological relevance of the candidates short listed from the *in vitro* experiments. Only a few *in vivo* reports that assign specific substrates to plant kinases such as the identification of ACS as an MPK6 substrate and of MKS1 as an MPK4 substrate in *Arabidopsis* are currently available. Compared to the status in mammals, we know very little about kinase substrates and their phosphorylation sites in plants. Therefore it is “time to search for protein kinase substrates” in plants.

33.5 FIVE-YEAR VIEWPOINT

Application of available phosphoproteomics technologies in plants is necessary to establish economic and reproducible procedures for the detection of phosphorylated proteins, the identification of their phosphorylation site(s), and the search for novel kinase substrates. Improved computational prediction as well as application of *in vitro* methods, such as protein-microarray based technologies, will speed up the efficient selection of candidate substrates prior to *in vivo* studies. Proteome arrays containing the complete proteome of a plant on microscopic slides are envisaged as an excellent platform in search for novel kinase substrates [32]. The increasing provision of plant protein expression clones as well as the progress in the field of cell free protein expression will provide comprehensive sets of plant proteins therewith easing the application of protein microarray-based technologies. The radioactive detection of phosphorylation on microarrays will be more and more replaced by sensitive fluorescence phospho dyes, such as Pro-Q DPS. The availability of large collections of recombinant proteins will help to generate plant protein specific antibodies. These antibodies can be applied for the enrichment of specific plant proteins from plant extracts and facilitate the *in vivo* detection of phosphorylated plant proteins. AMAs and RPMAs will become very attractive tools to monitor the phosphorylation state of plant proteins with high sensitivity *in vivo*, allowing for analyzing large numbers of samples with minute sample volumes. Their future application in the plant field will depend on the provision of high specific anti-phospho antibodies, especially anti-pSer and anti-pThr antibodies. Results on protein phosphorylation will feed into systems biology, where new hypotheses can be developed by computational approaches and then be proven using proteomic strategies.

REFERENCES

1. Stone JM and Walker JC. 1995. *Plant Physiol* **108**:451–457.
2. Mukherji M. 2005. *Expert Rev Proteomics* **2**:117–128.
3. Hardie DG. 1999. *Annu Rev Plant Physiol Plant Mol Biol* **50**:97–131.
4. Peck SC. 2006. *Plant J* **45**:512–522.
5. Manning G, et al. 2002. *Science* **298**:1912–1934.
6. Venter JC, et al. 2001. *Science* **291**:1304–1351.
7. Kersten B, et al. 2006. *Proteomics* **6**:5517–5528.
8. Berwick DC and Tavaré JM. 2004. *Trends Biochem Sci* **29**:227–232.
9. Yoshioka K, et al. 2004. *Mol Cells* **17**:223–229.
10. Liu Y and Zhang S. 2004. *Plant Cell* **16**:3386–3399.
11. Fields S and Song O. 1989. *Nature* **340**:245–246.
12. Soellick TR and Uhrig JF. 2001. *Genome Biol* **2**:RESEARCH0052.
13. Kersten B, et al. 2002. *Plant Mol Biol* **48**:133–141.
14. Kersten B, et al. 2004. *Curr Proteomics* **1**:131–144.
15. Andreasson E, et al. 2005. *EMBO J* **24**:2579–2589.
16. Menke FL, et al. 2005. *Mol Plant Microbe Interact* **18**:1027–1034.
17. Zhu H, et al. 2001. *Science* **293**:2101–2105.
18. Angenendt P, et al. 2006. *Mol Cell Proteomics* **5**:1658–1666.
19. Ramachandran N, et al. 2004. *Science* **305**:86–90.
20. Grelle G, et al. 2006. *Mol Cell Proteomics* **5**:234–244.
21. Ptacek J, et al. 2005. *Nature* **438**:679–684.
22. Khan M, Takasaki H, and Komatsu S. 2005. *J Proteome Res* **4**:1592–1599.
23. Asano T, et al. 2002. *Plant Cell* **14**:619–628.
24. Shah K, et al. 1997. *Proc Natl Acad Sci USA* **94**:3565–3570.
25. Eblen ST, et al. 2003. *J Biol Chem* **278**:14926–14935.
26. Fukunaga R and Hunter T. 1997. *EMBO J* **16**:1921–1933.
27. Fukunaga R and Hunter T. 2004. *Methods Mol Biol* **250**:211–236.
28. Bertone P and Snyder M. 2005. *FEBS J* **272**:5400–5411.
29. Feilner T, et al. 2004. *Curr Proteomics* **1**:283–295.
30. Hultschig C, et al. 2006. *Curr Opin Chem Biol* **10**:4–10.
31. Kersten B, et al. 2005. *Expert Rev Proteomics* **2**:499–510.
32. Kung LA and Snyder M. 2006. *Nat Rev Mol Cell Biol* **7**:617–622.
33. Kreutzberger J. 2006. *Appl Microbiol Biotechnol* **70**:383–390.
34. Kersten B, et al. 2004. *Anal Biochem* **331**:303–313.
35. Huang J, et al. 2004. *Proc Natl Acad Sci USA* **101**:16594–16599.
36. Kersten B, et al. 2003. *Plant Mol Biol* **52**:999–1010.
37. MacBeath G and Schreiber SL. 2000. *Science* **289**:1760–1763.
38. Zhu H, et al. 2000. *Nat Genet* **26**:283–289.

39. Feilner T, et al. 2005. *Mol Cell Proteomics* **4**:1558–1568.
40. Kramer A, et al. 2004. *Phytochemistry* **65**:1777–1784.
41. Feilner T and Kersten B. 2006. *Methods Mol Biol* **355**:379–390.
42. Kersten B and Feilner T. 2007. High-throughput identification of potential plant protein kinase substrates, In: *Plant Proteomics* (Samaj J and Thelen J, eds.), Springer, Berlin, pp. 53–70.
43. Martin K, et al. 2003. *Proteomics* **3**:1244–1255.
44. Pal M, et al. 2006. *Anal Chem* **78**:702–710.
45. Schutkowski M, Reineke U and Reimer U. 2005. *ChemBioChem* **6**:513–521.
46. Loog M, et al. 2000. *Eur J Biochem* **267**:337–343.
47. Nielsen UB, et al. 2003. *Proc Natl Acad Sci USA* **100**:9330–9335.
48. Chan SM, et al. 2004. *Nat Med* **10**:1390–1396.
49. Asai T, et al. 2002. *Nature* **415**:977–983.
50. Yang SH, Sharrocks AD, and Whitmarsh AJ. 2003. *Gene* **320**:3–21.
51. MAPK Group. 2002. *Trends Plant Sci* **7**:301–308.
52. Agrawal GK, Iwahashi H, and Rakwal R. 2003. *Biochem Biophys Res Commu* **302**: 171–180.
53. Zwerger K and Hirt H. 2001. *Biol Chem* **382**:1123–1131.
54. de la Fuente van Bentem S, et al. 2006. *Nucleic Acids Res* **34**:3267–3278.

TYROSINE PHOSPHORYLATION IN PLANTS: EMERGING EVIDENCE

Andrea Carpi, Valeria Rossi, and Francesco Filippini

34.1 INTRODUCTION

Mechanism and Relevance of Tyrosine Phosphorylation in Eukaryotes

Addition or removal of phosphate groups to/from amino acid residues can induce allosteric modifications resulting in conformational changes in proteins leading to either their activation or inactivation, or creating docking sites for other proteins and causing intracellular relocation. Reversible phosphorylation of proteins is one of the most common PTM mechanisms for regulating protein activity in various processes such as cellular growth, cell cycle control, cytoskeleton assembly, ionic current modulation, and receptor regulation. In eukaryotic cells, protein phosphorylation mainly concerns Ser, Thr, or Tyr residues of protein moieties. In typical animal cells, pTyr represents only a tiny fraction of total protein phosphorylation. However, Tyr phosphorylation plays a pivotal role in a variety of important signaling pathways in multicellular organisms. A constitutively activated Tyr kinase oncogene product can increase pTyr percentage up to tenfold, and this relatively minor change is sufficient to induce malignant transformation of the cell.

Enzymes Involved in Tyr Phosphorylation Pathways

The reversibility of protein Tyr phosphorylation relies on the coordinated actions of protein Tyr kinases (PTKs) and PTPs, both of which exist as large protein families. When considering whole proteomic kinase complements, PTKs represent a minor fraction of varying representativeness (e.g., ~5% and >15% in, respectively, *H. sapiens* [1] and *C. elegans* [2]).

PTKs and PTPs play a key role in cell signaling. RTKs are all type I TM proteins with a cytoplasmic domain that has intrinsic catalytic activity that is activated upon ligand binding. Ligand-induced dimerization of RTKs results in juxtaposing the two catalytic domains, allowing mutual transphosphorylation of residues in the activation loop of the catalytic domain. This process leads to kinase activation and autophosphorylation of Tyr residues outside the catalytic domain, which are the key to downstream signaling. Receptors lacking intrinsic catalytic activity can be coupled to nonreceptor PTKs (nrPTKs) via noncovalent association with the cytoplasmic domain of a receptor subunit, thus forming “binary” receptors.

Although phosphotransfer reactions catalyzed by various PTKs show similar basic mechanisms, their biological functions demonstrate a considerable degree of specificity. Catalytic domains of pPTKs can differentiate between Tyr(s) in a common physiological substrate. Each kinase produces a distinct pattern of Tyr phosphorylation, supporting the concept that catalytic domains determine the final sites of phosphorylation once PTKs reach their target proteins. Protein substrates of PTKs are often difficult to discern, but recent methods have been of help to identify targets and characterize their structural interactions with kinases. X-ray analysis of PTK-substrate co-crystal structures has revealed both conserved and specialized features of recognition, which probably contribute to substrate selection and the individual functions of these key enzymes [3].

In addition to PTKs, further protein kinases can mediate Tyr phosphorylation; these enzymes can phosphorylate also Ser/Thr residues and are therefore referred to as “dual specificity” kinases (DSKs). DSKs that phosphorylate the Thr and Tyr residues within the T-x-Y motif of MAPKs play a pivotal role in the control of various processes of cell growth. These DSKs—also known as MAPKKs—are rings of a sequential kinase signaling chain [4]. A further DSK family is represented by dual-specificity Tyr-regulated kinases [5] that are highly conserved mediators of growth control and differentiation.

The function of PTPs is not simply to scavenge pTyr, and the past decade has uncovered a wide range of signaling pathways that are regulated by PTPs. These enzymes have distinct but complementary function, because kinases have been implicated in controlling the amplitude of a signaling response, whereas phosphatases seem to have an important role in regulating the rate and duration of the response [6]. Unlike the protein kinases, which are derived from a common ancestor, the protein phosphatases have evolved in separate families that are structurally and mechanistically distinct. PTPs are encoded by the largest family of phosphatase genes. Phosphate is removed from pTyr proteins by three families of Cys-based PTPs, and a family of metal-ion-dependent Asp-based PTPs [7]. Cys-based PTPs are defined by the

active-site signature motif H-C-x(5)-R, in which the Cys residue functions as a nucleophile and is essential for catalysis [6]. The largest family of Cys-based PTPs is related to PTP1B, a nonreceptor PTP (nrPTP), and to CD45, a receptor-like PTP (RPTP). Within this family are also some PTPs with predominantly nonprotein substrates, including phospholipids and RNA. Enzymes referred to as the “classical” PTPs have a deep catalytic pocket, which only allows the long side chain of pTyr to reach the catalytic machinery; thus, these PTPs are strictly Tyr-specific. Other “dual specificity” phosphatases (DSPs) have a more shallow catalytic pocket, and many of them dephosphorylate both pSer/pThr and pTyr. These include the MAP kinase phosphatases (MKPs) that dephosphorylate the mitogen-activated protein kinases at their dually phosphorylated Thr-x-Tyr activation loop motif. Another subgroup of “atypical” DSPs includes a number of low- M_r enzymes (<250 amino acid residues) lacking specific targeting motifs. Several members of the PTP superfamily possess conserved domains with core features of a PTP, but which lack residues that are critical for catalysis. Such “pseudophosphatase” PTPs are involved in interactions that regulate both the enzymatic activity and the subcellular location of the active phosphatase [6].

The architecture of the active site of members of the PTP superfamily renders these enzymes sensitive to reversible oxidation and inactivation; this mechanism plays an important role in controlling the signal output following an extracellular stimulus [6]. RPTPs are predominantly found in the PM, whereas nrPTPs are localized to a variety of intracellular compartments, including the cytosol, PM, and ER. Categorization as RPTPs or nrPTPs is not an absolute distinction, because use of alternative promoters or alternative splicing can lead to the production of TM and cytoplasmic forms of some PTPs from a single gene [6].

Relationship to Tissue and Organ Development

Over the past two decades, it has become clear that Tyr phosphorylation plays a pivotal role in a variety of important signaling pathways in multicellular organisms. It is deeply involved in the development of multicellular organisms, including higher eukaryotes such as mammals as well as lower eukaryotes such as sponge and hydra [8] and even prokaryotes such as *M. xanthus* [9]. In animals, PTKs and PTPs play important roles in the regulation of cell growth and differentiation [10]. Because unregulated Tyr phosphorylation signaling causes a breakdown in the normal regulation of cell proliferation and motility, leading to diseases including cancer, this signaling pathways is now a major focus of biomedical research.

PhosphoTyr-Based Signaling (SH2, etc)

The discovery of the SH2 domain recognizing specific pTyr residues in activated RTKs and their targets revealed how activated PTKs propagate signals and the association of two proteins could be induced by phosphorylation thus propagating a signal [11]. This illustrated a new function for protein phosphorylation, namely the regulation of protein–protein association. SH2 domains bind in a sequence-specific fashion,

recognizing residues C-terminal to pTyr. After SH2, further PTB domains have been found. PTB domains recognize residues N-terminal to pTyr, but only a subset of PTB domains binds to their target proteins in a phosphorylation-dependent manner [11]. SH2 and PTB domain interactions mediate the recruitment of target proteins to activated PTKs—thus allowing for their phosphorylation—and the translocation to the PM. SH2 and PTB domains can be present in proteins endowed with enzymatic activity as well as in so-called adaptor proteins (e.g., Grb2) that bind to and thereby bring effector enzymes to the PM. In addition to SH2 and PTB domains, novel pTyr-binding domains have recently been identified. A Gab1 pTyr-binding domain appears to have a novel structure; a further type of pTyr-binding domain is shown by “anti-phosphatases.” These anti-phosphatases are structurally related to either PTPs or DSPs, but lack residues crucial to catalytic activity and therefore can bind but not hydrolyze pTyr-containing proteins [12].

34.2 BRIEF BIBLIOGRAPHIC REVIEW

Protein Phosphorylation in Plants

Plant proteomes contain a high number of eukaryotic protein kinases and phosphatases [13, 14] and a few histidine kinases, related to bacterial two-component systems [15]. Several plant Ser/Thr kinases belong to the classical groups that are found also in animals; in addition, plants are endowed with a family of RL(Ser/Thr)Ks that are topologically similar to mammalian RTKs and are treated elsewhere in this book. A robust body of evidence supports a relevant role for Ser/Thr-specific kinases in the regulation of plant growth, development, and physiological responses. Instead, even the existence of Tyr phosphorylation in plants has been controversial until recently, when several PTPs and a number of DSPs were characterized [6, 14, 16].

Early Evidence on Tyr Phosphorylation in Plants

In the first half of the 1990s, early evidence on Tyr phosphorylation in plants was reported. Cloning, expression in *E. coli*, and biochemical characterization of the product of *A. thaliana* APK1 gene allowed authors to conclude that this novel type of protein kinase could phosphorylate Tyr, Ser, and Thr residues, though Tyr phosphorylation seemed to occur only on limited substrates [17]. Then, further evidence was obtained for the presence in plants of DSKs [18], and the dual phosphorylation of the Thr-x-Tyr motif in the activation loop of MAPKs was found to be conserved also in plants [19]. The first example of Tyr phosphorylation in plant organelles also represented the first indication of protein phosphorylation as part of a redox signaling mechanism in mitochondria [20].

Involvement of Tyr Phosphorylation in Light-Regulated Signaling

In subsequent years, further evidence on the presence of PTK activity in plants was obtained. The characterization of a PTP encoded by a stress-responsive gene

in *Arabidopsis* [21] and evidence that in plants dual-specificity MAPKK are clearly involved in responses to biotic and abiotic stress [4] suggested that reversible Tyr phosphorylation of a limited number of plant proteins might mediate stress responses [21]. However, pTyr was found to be 0.5% of the total phosphoamino acids labeled with [32 P]orthophosphate in endogenous maize seedlings proteins, and biochemical and immunological data indicated the presence of Tyr kinase activity and also pTyr in proteins of maize seedlings [22]. Using anti-pTyr antibodies as well as substrates and inhibitors specific to PTKs, further research groups could demonstrate that Tyr phosphorylation of plant proteins is likely to regulate—in addition to stress response—most important pathways, growth, and development. Tullberg et al. [23] found that Tyr phosphorylation of plant organellar proteins also concerns chloroplasts; in particular, anti-pTyr antibodies reacted specifically with several thylakoid proteins, two of which were components of the LHC II, likely involved in the regulation of photosynthetic electron transfer. Then, the PTK inhibitor genistein, at concentrations causing no direct effect on Thr kinase activity, was found to prevent Tyr phosphorylation of LHC II, the transition to light state 2, and associated Thr phosphorylation of LHC II [24]. Involvement of Tyr phosphorylation in light-regulated signaling processes was further confirmed by functional evidence on a DSK targeted to the chloroplast in tobacco [25].

Regulation by Tyr Phosphorylation of Plant Bending and Stomatal Opening

In year 2000, protein Tyr phosphorylation was reported to be involved in plant bending [26]. The author's [26] performed 2D-PAGE and immunodetection experiments with anti-actin and anti-pTyr antibodies, and they also used phenylarsine oxide (PAO), a specific inhibitor of PTPs that is known to inhibit Tyr dephosphorylation of actin hence cell motility. Those experiments allowed them to demonstrate that actin—in the contact-sensitive plant *M. pudica*—is heavily Tyr-phosphorylated and that changes in the extent of phosphorylation correlate with the degree of bending of the plant's petioles. Since actin is dynamically arranged in guard cells, Tyr phosphorylation was proposed to be likely involved also in the regulation of the opening and closing of these cells [26]. Indeed, this was confirmed by functional evidence. Using several specific PTP inhibitors, it was demonstrated that PTP activity is essential for stomatal closure induced by four different factors including ABA, external calcium, darkness, and H₂O₂ [27]. This study identified efflux from vacuole, but not fluxes across the PM, as the target process for the PTP inhibitors, suggesting that a protein specifically related to tonoplast K⁺ channels is regulated by reversible Tyr phosphorylation. Thus, it is conceivable that Tyr phosphorylation of actin is somehow related to the regulation of K⁺ fluxes in motor cells [28]. Evidence from these studies and the characterization of several PTPs from *Arabidopsis* and other species has been able to modify the earlier tenets [21] on Tyr phosphorylation in plant cell signaling and regulation.

Tyr Phosphorylation in Plants Depends on Developmental Conditions

We can agree with Luan [28] that a hard blow to the paradigm that Tyr phosphorylation is not important in plant cells probably came from finding that MAPK cascades are responsible for transducing numerous biotic and abiotic stress and hormonal signals in plants [4, 29]. Moreover, further reports showed that Tyr phosphorylation is deeply involved also in plant development. Evidence was obtained on the occurrence of Tyr kinase activity in developing coconut (*C. nucifera*) zygotic embryos, based on *in vitro* phosphorylation with radiolabeled ATP, alkaline treatment, and thin-layer chromatography analysis [30]. Tyr phosphorylation was further confirmed in extracts of embryos at different developmental stages by anti-pTyr immunoblotting that revealed a changing profile of pTyr-proteins during embryo development [30]. Using RR-SRC, a PTK-specific synthetic peptide derived from the amino acid sequence surrounding the phosphorylation site in pp60(src), Tyr kinase activity was found to show two peaks of activity, during early and late embryo development [30]. The possibility that Tyr phosphorylation may play a role in the coordination of plant embryo development was further supported by analysis of carrot (*D. carota*) cell cultures, which represent a model system for somatic embryogenesis. Recombinant anti-pTyr antibodies allowed authors to identify a large number of pTyr proteins from all subcellular fractions; recognition specificity was assessed by displacement experiments as well as by on-blot Tyr dephosphorylation [31]. Patterns of pTyr-proteins were found to vary among different adult plant tissues or somatic embryo stages, and somatic embryogenesis was blocked *in vivo* by a cell-permeable tyrphostin (a specific inhibitor of PTKs), demonstrating that Tyr phosphorylation depends on the developmental conditions and is involved in the control of specific steps in plant development. Developmental regulation of Tyr phosphorylation was demonstrated also in peanut (*A. hypogaea*): A STY kinase was found to be involved in abiotic stress (cold and salt) responses and in seed development [32]. Such peanut STY is a structural mosaic of Ser/Thr and Tyr kinases that autophosphorylates predominantly on Tyr, and it represents a prototypical member of a new division in the plant kinase family, as well as the first report of a non-MAP kinase cascade DSK involved in abiotic stresses.

Reversible Tyr Phosphorylation Mediates Responses to Phytohormones

Phytohormone-stimulated cell cycle reactivation in hypocotyls of *A. thaliana* was accompanied by Tyr phosphorylation of several proteins. This was not observed in a callus formation-deficient mutant, suggesting that the induction of Tyr phosphorylation occurs as a specific event that may play an important regulatory role in phytohormone stimulated cell proliferation [33]. The action of cytokinin in stimulating the activation of CDK by removal of inhibitory phosphorylation from Tyr is the key primary effect of the hormone in its stimulation of cell proliferation, since cytokinin could be replaced by expression of *cdc25*, which encodes the Cdc25 LMW PTP [34]. Thus, Tyr dephosphorylation in CDK is a normal part of prophase progression in plants and a positive biochemical signal of cytokinin presence, stimulating cell proliferation and

plant development. 2-DGE followed by immunoblotting with monoclonal anti-pTyr antibodies revealed BR-induced Tyr phosphorylation of several proteins in pea leaves, possibly including subunits of RuBisCO [35]. Tyr dephosphorylation was found to play an important role also in the control of post-germination arrest of development by ABA in *A. thaliana* [36]. Inhibition of Tyr dephosphorylation by PAO in the presence of ABA results in a delay in seed germination depending on increased ABA sensitivity, accompanied by hyper-induction of the ABA-responsive genes [36].

34.3 METHODOLOGY AND STRATEGY

Asynchronous Flow of Functional Evidence and Proteome-Wide Analyses

In this “omics” era, signaling molecule families are being defined and the complexities of the control of signal transduction are now becoming apparent. Genome projects are currently producing many thousands of gene sequences from which protein amino acid sequences may be deduced. The deposition of such data in public databases before a paper can be published allows proteome bioinformaticians to perform *in silico* analyses based on very large data sets, increasing prediction confidence and inference power that were previously impaired by either unbalanced organism source or poor data sets. Experimental research on different organisms and taxonomic lineages has provided functional evidence concerning a high number of genes and proteins, as well as molecular and cellular pathways. Most often, the characterization of genes and proteins from different organisms is asynchronous. For instance, when considering two genes “X” and “Y,” it may happen that the function of gene “X” may be unraveled by the characterization of a *Drosophila* mutant, while the enzymatic activity of protein encoded by gene “Y” of *Arabidopsis* is reported. Although functional evidence is not simply forwardable from a gene/protein to the orthologous genes or proteins, the asynchronous flow of discoveries has already allowed several scientists to perform comparative analyses aimed at enhancing the elucidation of important mechanisms. Thus, computational analysis can be exploited to fuel hypothesis-driven experimental research through the exploration of online databases; and in the last decade of the 20th century as well as in recent years, a progressively increasing number of genome/proteome-wide comparative analyses resulted in both defining novel sequence marks and identifying genomic/proteomic complements.

Identification of Putative Plant PTKs by Comparative Proteome Bioinformatics

Most of the >900 putative protein kinases detected *in silico* in the proteome of *A. thaliana* have been tentatively assigned to the Ser/Thr class [37]. However, putative specificity reported in annotated sequences is commonly inferred by similarity from other sequences; and this is inappropriate in the case of eukaryotic protein kinases, which are known to share a highly conserved catalytic domain while residue and substrate phosphorylation specificities are determined by motif variation and by

noncatalytic regions. Given that experimental evidence on phosphorylation specificity concerns only a limited subset of protein kinases from a given proteome, and considering that in animals PTKs represent a minor fraction of the kinase complement, the presence of PTKs and/or PTK-like kinases among as yet uncharacterized products of plant kinase genes cannot be excluded *a priori*.

Although no classical PTK has hitherto been cloned from plants, a comparative bioinformatic analysis identified a complement of putative PTKs in *A. thaliana* [38]. Such putative kinases were identified by screening the whole proteome with highly specific sequence markers. In fact, the protein kinase domain is too conserved among Ser/Thr and Tyr kinases to allow specific detection of putative PTKs by homology search. This problem was overcome using as sequence probes patterns PS00108 and PS00109 from the well-known database PROSITE, which contains “similarity-independent” amino acid signatures. In order to further improve the extraction of putative PTKs, Carpi et al. [38] took into account the problem of true hits, endowed with the proband activity but showing a degenerate signature (FN hits) and screened the *Arabidopsis* proteome using also slightly degenerate PROSITE signatures as sequence probes. Positive hits showing either canonical or degenerate patterns were then analyzed to discard FP hits corresponding to nonkinase proteins or nonfunctional kinases (e.g., lacking one or more subdomains, or mutated at residues crucial to the enzyme activity). Then, all putative PTK or DSK sequences were further analyzed to group them based on conserved or specific moieties, and the expression of corresponding genes in green tissues was assessed. Strategy that led to the identification of a complement of putative PTKs and DSKs in *Arabidopsis* is summarized in Figure 34.1.

Genome-Wide and Tyr Phosphoproteome Analysis of Plant STY Kinases

In order to address the question whether *bona fide* PTKs exist in plants, a genomic survey of protein Tyr kinase motifs in plants was performed using delineated Tyr phosphorylation motifs from the animal system [39]. In particular, plant proteomes were scanned to identify plant sequences containing a consensus that is conserved in subdomain XI of a number of biochemically characterized PTKs from mammals, *Drosophila* and *Dictyostelium*. After elimination of redundancy and FPs, sequences were scanned for the presence of motifs specific to other protein kinase subdomains; the results of this scanning (see *Bioinformatic Identification of Proteomic Complements*) suggested naming identified sequences as STY kinases [39]. Phylogenetic and domain architecture analysis allowed authors to cluster STY kinases into families and to infer possible involvement in signaling pathways, while hydropathy plots were used to predict subcellular localization. Finally, homology modeling of the STY kinases was performed, and on-line data from *Arabidopsis* microarray experiments allowed for grouping of STY genes based on their expression during a specific growth stage in a particular organ or following an environmental stimulus [39].

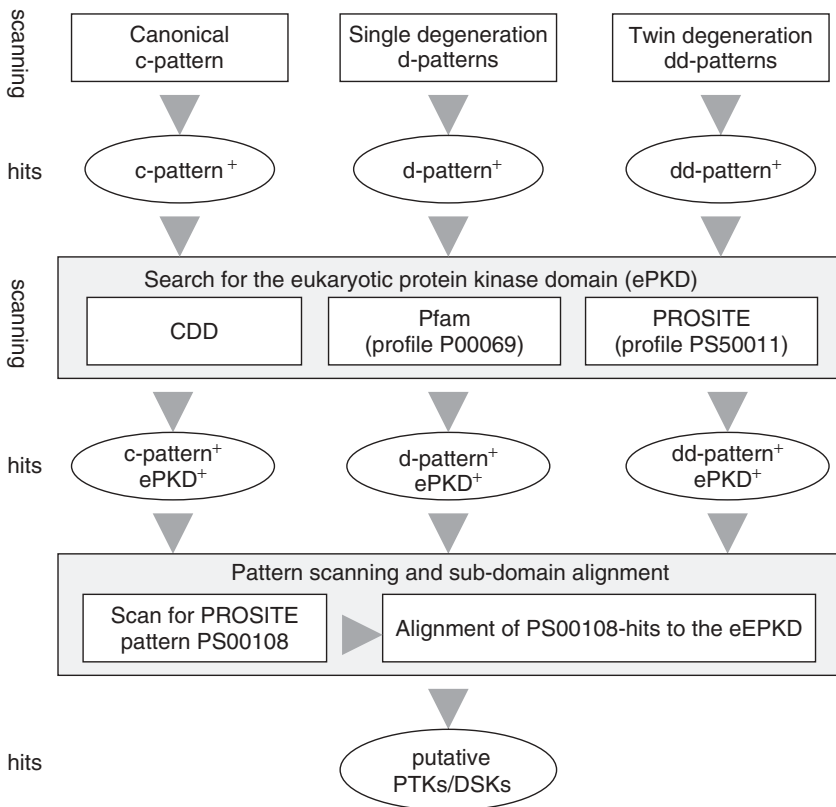


FIGURE 34.1. Summary of the strategy followed by Carpi et al. [38] to identify *in silico* a complement of putative PTKs and DSKs in *A. thaliana*. A whole proteome scanning was performed to search for protein sequences having a canonical PROSITE pattern PS00109 for PTKs (c-pattern) or a degenerate version of PS00109 characterized by one (d-pattern) or two (dd-pattern) noncanonical residues. Positive hit sequences (c-pattern⁺, d-pattern⁺, dd-pattern⁺) were further analyzed and forwarded to next steps when showing a profile for the eukaryotic protein kinase domain (ePKD⁺). Sequences found to have a PS00108 pattern in subdomain VIIb of the catalytic domain were discarded, because this signature is highly specific to Ser/Thr kinases. Finally, each sequence was aligned to the consensus of the ePKD for checking the presence of subdomains and conserved residues.

***In Silico* Identification of Plant pTyr Phosphatases**

In addition to kinases, both phosphatases and signaling domains have been investigated by *in silico* proteomics. Structural sequence probes from aligned and unaligned sequence models, and all-against-all BLAST searches were used to identify the phosphatase catalytic subunit sequences in the *Arabidopsis* proteome; then, phylogenetic analyses were used to infer clustering patterns reflecting sequence similarity and evolutionary ancestry [40]. This study led to the identification of a number of DSPs, many of them with no obvious clues to function. Working independently, Fordham-Skelton

et al. [41] discovered the sequences of a small set of DSPs, and they showed that these sequences contain a nonphosphatase domain (kinase interaction sequence, KIS). Further bioinformatic investigations, including homology search and domain profile scanning, demonstrated that this nonphosphatase domain is phylogenetically widespread and encodes an ancient carbohydrate-binding domain [42].

***In Silico* Identification of pTyr Binding SH2 Domains in Plants**

PTB protein domains are central players in signal transduction pathways mediated by reversible Tyr phosphorylation. Considering evidence on plant protein Tyr phosphorylation and the identification of complements of putative PTKs, PTPs, DSKs and DSPs, plant proteomes were scanned for the presence of SH2 domains. A BLAST search was performed using the SH2 domain of *Dictyostelium* STAT (signal transducer and activator of transcription) protein as the query sequence; after identification of plant SH2 domain-containing proteins and multiple alignment analysis, the 3D structure of the plant SH2 domains was predicted by homology modeling [43]. Gaps in the target sequence were subjected to local energy minimization to bring the core ends together and to alleviate local conformational strain; insertions in the target sequence were modeled by searching a fragment database of high-resolution structures (<1.5 Å) to find an appropriate template [43]. Secondary structural elements in the model were found to fold into what should constitute a functional PTB module.

34.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Bioinformatic Identification of Proteomic Complements

Scanning of the *Arabidopsis* proteome for the presence of pattern PS00109 (see *Identification of Putative Plant PTKs by Comparative Proteome Bioinformatics* and Figure 34.1) and subsequent elimination of FPs suggested that PTKs might represent roughly 3% of the kinase complement in plants [38]. These putative kinases were clustered into three major groups based upon homology, domain architecture, and predicted subcellular localization. Group 1 kinases fall in the RLK-Pelle group, although they are putatively endowed with Tyr-specific activity; these kinases can be further divided in three subgroups: (i) putative RTKs, sharing—in most members—LRR in the predicted extracellular domain; (ii) putative cytoplasmic PTKs related to Pti1, a downstream interactor of the Pto kinase (a plant homologue of *Drosophila* Pelle), and (iii) putative PTKs related to wall-associated kinases (WAK). Group 2 kinases belong to the RIO1 family, while kinases in Group 3 are similar to MAPK kinases or eIF 2 α protein kinases, which are both endowed with DSK activity [38]. It is noteworthy that kinases in the RLK-Pelle superfamily play a major role in the control of organ and tissue development as well as in the response to pathogens, and WAKs are expressed at organ junctions, in shoot and root apical meristems, in expanding leaves [38]. A more recent bioinformatics analysis [39]—once again performed using *Arabidopsis* as a model plant—has led to the identification of several members of the STY class

of protein kinases that seem to have cytoplasmic localization. STY kinases are structurally and functionally different from both PTKs and *sensu stricto* DSKs and seem to be regulated by Tyr phosphorylation [44].

Recall and Specificity in Pattern Scannings

It is not surprising that when scanning the same proteome (*Arabidopsis*) by two different patterns—although they are both specific to Tyr kinases—quite different complements of putative PTKs/DSKs are identified. In fact, both the PROSITE pattern PS00109 of subdomain VIb [38] and the CW-x(6)-RP-x-F pattern of subdomain XI [39] have been originally derived by multiple alignment of sequences showing unbalanced taxonomic distribution. This is not uncommon: Most patterns and profiles are derived from preliminary, narrow sets of sequences and hence show a reduced recall index (several FNs are not recognized). Then, patterns are usually updated to improve both recall (by accepting residues specific to FNs) and precision (by excluding those residues that are present in FPs but not in TP hits). Thus, when using a signature common to a taxonomic group to scan the genome/proteome of evolutionarily distant organisms, degeneration has to be taken into account. Moreover, the catalytic activity of a protein kinase depends on the whole complex of its subdomains rather than on a specific subdomain, as demonstrated by evidence that FP and FN hits are associated to most patterns.

The canonical CW-x(6)-RP-x-F of subdomain XI is slightly degenerate in putative PTKs from *Arabidopsis* [38]. However, when screening public databases with degenerate CW-x(6)-RP-x-F patterns, we found that also animal PTKs and DSKs may have a degenerate motif at subdomain XI. After having discarded degenerate motifs from the second catalytically inactive domain of some kinases, we found that degeneration can be apparent also in active kinases endowed with a single catalytic domain. Position 2 is Trp (aromatic) in the canonical pattern, but it can be either Ser (polar) in mouse CSK (a PTK) or Leu/Met (aliphatic) in DSKs MAPKK6 and MAPKK7, or even Cys in testis-specific mammalian DSKs. Position 12 (Phe) can be degenerate (Leu or Tyr) in both cytoplasmic (mouse Tec) and receptor-like (*Xenopus* and chicken Ephrin type-B) PTKs as well as in *Drosophila* MAPKK Q23977.

Putative PTKs in *O. sativa* and Other Plant Species

We report here that putative PTK and DSK sequences [38] are present also in the proteome of rice and of other plant species. Pattern scanning (using both canonical and degenerate PS00108/PS00109 patterns as sequence probes) was followed by homology search and by domain architecture analysis. The relative representativeness of putative PTKs in rice was found to be lower than in *Arabidopsis* (roughly 2.5% of the protein kinase proteomic complement). This can be attributed to the amplification of the Ser/Thr kinase family, because the number of putative PTKs in rice is even (slightly) higher than in *Arabidopsis*. Homology-based analysis resulted in clustering of rice putative PTKs into the same groups as *Arabidopsis* [38], but the number of members within each group/subgroup is different between the two species. The rice proteome

contains a higher number of putative PTKs that are predicted to have receptor-like topology; instead, Pti1-like proteins are more represented in *Arabidopsis*. Comparison of the two complements of putative PTKs, aimed at identifying possible orthologues and paralogues, provided intriguing evidence. Pairs of orthologues represent only 50% (or even less) of the sequences in the two sets; the analysis of the other sequences suggests that differential amplification of paralogues might have occurred in the two species (and/or in their ancestors). Neither orthologues nor paralogues can be identified when using as sequence probes a number of putative PTKs from both *Arabidopsis* and rice, especially because of high divergence in noncatalytic regions. Last but not least, there are (i) a few pairs showing highly similar sequences but different motifs (e.g., PS00109 in *Arabidopsis* and PS00108 in rice) and (ii) putative kinases that are predicted to be endowed with two catalytic domains, each showing a different motif. Canonical or degenerate PS00108/PS00109 patterns were found also in the sequences of putative kinases from plant species other than *Arabidopsis* and rice; however, incomplete genome sequencing renders it possible to infer the presence, but not the absence, of members in specific kinase groups.

34.5 CONCLUSIONS

More than 10 years of investigation on Tyr phosphorylation in plants have produced a robust body of evidence concerning Tyr phosphorylation of a high number of plant proteins that are present in all subcellular fractions and are involved in most important physiological and developmental pathways. These results are in agreement with evidence that main protein families and pathways common to multicellular eukaryotes are shared also by *A. thaliana* [37]. After sequencing the genome of this model plant, further evidence could be obtained by comparative proteome bioinformatics. Several groups, working independently, have identified a considerable number of “players” (single proteins, groups, or even whole complements) of the Tyr phosphorylation system: putative PTKs, STYs, DSKs, PTPs, DSPs, and SH2 domain containing proteins. Presence of some FPs among such sequences cannot be excluded; however, multiple signatures, motifs, profiles, and folds that are characteristic of the Tyr phosphorylation system are conserved also in a relevant number of proteins from various plant species. This is in perfect agreement with biochemical, cellular, and physiological evidence already supporting both existence and role of enzymes mediating reversible Tyr phosphorylation of plant proteins: DSKs from various classes, as well as PTPs and DSPs. It remains to be seen whether plants possess *sensu stricto* PTKs. None has been characterized so far, but a relatively few, key amino acid differences distinguish the kinase domains of PTKs from those of DSKs and Ser/Thr kinases, suggesting that they might have been overlooked. The subdomain motifs so far identified might have varying relevance to the control of phosphorylation specificity in animal and plant protein kinases. It cannot be excluded that protocols to characterize animal PTKs might be suboptimal or even nonoptimal to characterize their plant counterparts. Moreover, it is tempting to speculate that, in agreement with high “plasticity” shown by plants at genomic,

cellular, and organism level, control of specificity might be more “relaxed” in plant kinase domains. Thus, either dual specificity or prevalence of Tyr versus Ser/Thr specificity might depend on further (and as yet unknown) regulatory factors and conditions.

34.6 FIVE-YEAR VIEWPOINT

Presently, phosphorylation is the best understood PTM of proteins. However, the analysis of phosphoproteins and phosphopeptides is still one of the most challenging tasks in current proteome research. The body of experimental evidence obtained so far strongly indicates that many plant proteins from all subcellular compartments undergo phosphorylation on Tyr residues, but only a limited number of DSKs, PTPs and DSPs have been already characterized, and phosphorylation specificity of putative PTKs identified *in silico* is still waiting for confirmation. Only in a few cases, DSK(s), substrate(s), and PTP/DSP(s) could be attributed to the same pathway. Thus, further experimental work will need to shed light on major players and complete signaling cascades in the Tyr/dual phosphorylation system. Since not every phosphoprotein is accessible by a certain method and identification of the phosphorylated amino acid residue is required in the majority of cases, novel and efficient tools and strategies are being actively developed to allow, beyond the primary identification of phosphorylated proteins, the identification of phosphorylation sites and ultimately their quantification. These approaches are being used at various scales, from functional analyses to cataloguing studies, and current developments are likely to mark the transition towards the introduction of phosphoproteomics as one of the main integration levels in post-genome plant biology. Reduction of the sample complexity is one major step for the analysis of low-abundance kinase substrates (e.g., probably several PTK/DSK substrates); this can be achieved by enrichment of phosphorylated proteins or peptides by immunoprecipitation or MOAC, prior to analysis. Recently, developed phosphoproteomic approaches offer new possibilities for the identification of specific kinase targets. When considering that pTyr accounts for only 0.05% of the total cellular phospho-amino acid content (in animals) yet plays an unusually prominent role in signaling, development, and growth, tracking temporal and positional pTyr changes across the cellular proteome (i.e., Tyr phosphoproteomics) is tremendously valuable [45]. A nice example of Tyr phosphoproteomic approach has recently been described. A prototype antibody microarray platform is used to monitor changes in protein Tyr phosphorylation [45]. Different antibodies, each recognizing a specific cellular protein, are arrayed on a chip, incubated with total cell or tissue extracts or with biological fluids, and then probed with a fluorescently labeled, anti-pTyr monoclonal antibody. The protocol is optimized to allow for detection of changes in the Tyr phosphorylation state of selected proteins using submicrogram to low nanogram of total protein extract; the results are in agreement with previous, exhaustive MS analyses. Such antibody microarray method offers advantages of low sample and reagent consumption, scalability, detection multiplexing, and potential compatibility with microfluidic devices and automation.

REFERENCES

1. Robinson DR, Wu YM, and Lin SF. 2000. *Oncogene* **19**:5548–5557.
2. Plowman GD, et al. 1999. *Proc Natl Acad Sci USA* **96**:13603–13610.
3. Bose R, et al. 2006. *Curr Opin Struct Biol* **16**:668–675.
4. Mizoguchi T, Ichimura K, and Shinozaki K. 1997. *Trends Biotechnol* **15**:15–19.
5. Mercer SE and Friedman E. 2006. *Cell Biochem Biophys* **45**:303–315.
6. Tonks NK. 2006. *Nat Rev Mol Cell Biol* **7**:833–846.
7. Rayapureddi JP, et al. 2005. *Biochemistry* **44**:751–758.
8. Suga H, et al. 1999. *J Mol Evol* **8**:646–653.
9. Frasch SC and Dworkin M. 1996. *J Bacteriol* **178**:4084–4088.
10. Hunter T. 1995. *Cell* **80**:225–236.
11. Schlessinger J and Lemmon MA. 2003. *Sci STKE* **2003**(191):RE12.
12. Hunter T. 1998. *Nat Genet* **18**:303–305.
13. Stone J and Walker JC. 1995. *Plant Physiol* **108**:451–457.
14. Luan S. 2003. *Annu Rev Plant Biol* **54**:63–92.
15. Mizuno T. 2005. *Biosci Biotechnol Biochem* **69**:2263–2276.
16. Landrieu I, et al. 2004. *Proc Natl Acad Sci USA* **101**:13380–13385.
17. Hirayama T and Oka A. 1992. *Plant Mol Biol* **20**:653–662.
18. Ali N, Halfner U, and Chua NH. 1994. *J Biol Chem* **269**:31626–31629.
19. Usami S, et al. 1995. *Proc Natl Acad Sci USA* **92**:8660–8664.
20. Hakansson G and Allen JF. 1995. *FEBS Lett* **372**:238–242.
21. Xu Q, et al. 1998. *Plant Cell* **10**:849–857.
22. Trojanek J, et al. 1996. *Eur J Biochem* **235**:338–344.
23. Tullberg A, Hakansson G, and Race HL. 1998. *Biochem Biophys Res Commun* **250**:617–622.
24. Forsberg J and Allen JF. 2001. *Photosynthesis Res* **68**:71–79.
25. Cho HS, et al. 2001. *FEBS Lett* **497**:124–130.
26. Kameyama K, et al. 2000. *Nature* **407**:37.
27. MacRobbie EA. 2002. *Proc Natl Acad Sci USA* **99**:11963–11968.
28. Luan S. 2002. *Proc Natl Acad Sci USA* **99**:11567–11569.
29. Zhang S and Klessig DF. 2001. *Trends Plant Sci* **6**:520–527.
30. Islas-Flores I, Oropeza C, and Hernandez-Sotomayor SM. 1998. *Plant Physiol* **118**:257–263.
31. Barizza E, et al. 1999. *FEBS Lett* **447**:191–194.
32. Rudrabhatla P and Rajasekharan R. 2002. *Plant Physiol* **130**:380–390.
33. Huang HJ, et al. 2003. *Plant Cell Physiol* **44**:770–775.
34. Zhang K, Diederich L, and John PC. 2005. *Plant Physiol* **137**:308–316.
35. Fedina EO, Karimova FG, and Tarchevsky IA. 2006. *Biochemistry (Moscow)* **71**:423–429.
36. Reyes D, et al. 2006. *Planta* **223**:381–385.
37. The *Arabidopsis* Genome Initiative. 2000. *Nature* **408**:796–815.
38. Carpi A, et al. 2002. *Proteomics* **2**:1494–1503.

39. Rudrabhatla P, Reddy MM, and Rajasekharan R. 2006. *Plant Mol Biol* **60**:293–319.
40. Kerk D, et al. 2002. *Plant Physiol* **129**:908–925
41. Fordham-Skelton AP, et al. 2002. *Plant J* **29**:705–715.
42. Kerk D, et al. 2006. *Plant J* **46**:400–413.
43. Williams JG and Zvelebil M. 2004. *Trends Plant Sci* **9**:161–163.
44. Rudrabhatla P and Rajasekharan R. 2004. *Biochemistry* **43**:12123–12132.
45. Gembitsky DS, et al. 2004. *Mol Cell Proteomics* **3**:1102–1118.

14–3–3 PROTEINS: REGULATORS OF KEY CELLULAR FUNCTIONS

Peter C. Morris

35.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

14–3–3 proteins are dimeric proteins of approximately 30-kDa subunit size and a pI of 4–5, and they bind to partner proteins (also termed clients) over specific binding sites in both parties and regulate many aspects of cellular biochemistry such as enzyme activity, cell signaling, the cell cycle, protein targeting, and gene transcription. Because of their ubiquity in eukaryotes, their occurrence in almost all subcellular compartments, and their involvement in the regulation of many key metabolic and signaling pathways, they have become the subject of intense investigation and the topic of many recent excellent reviews [1–5]. 14–3–3s were first identified in 1967 from mammalian brain tissues through a combination of ion exchange chromatography and starch gel electrophoresis and earned their peculiar name from the column fraction number and gel coordinates [6]. The first recognized function of 14–3–3 in brain tissue was the activation of Try and Tyr monooxygenases [7]. Some time later, it was therefore surprising when what was thought to be a mammalian brain-specific protein was found by different groups of workers in spinach and *Oenothera* [8], in the model plant *Arabidopsis* and in maize [9, 10], and also in barley [11]. It soon became

apparent that 14-3-3s were to be found in all eukaryotic organisms in multigene families, varying in size from two in yeast [12] to seven in humans [13], eight in rice [14], and 13 expressed genes as well as two pseudogenes in *Arabidopsis* [15]. Additionally, the number of protein isoforms can vary due to mRNA splicing variants and/or PTMs such as phosphorylation or acetylation [1, 13]. Since 14-3-3 proteins had previously been fortuitously identified in different biological systems, they have historically been allocated alternative names, such as GF14 in *Arabidopsis*, MSF in rat, and the *rad24* gene in fission yeast. The core region of about 230 amino acids is well-conserved, but the short N- and C-termini are very divergent and may play a role in the specificity of binding. 14-3-3 has some amino acid sequence homology with the tetratricopeptide repeat proteins, which are also involved in the mediation of PPIs.

14-3-3 Structure

The crystal structure of mammalian 14-3-3 [16] has shown that the protein consists of nine antiparallel α -helices; dimerization (both hetero- and homodimerization is possible) occurs over reciprocal contact of helix 1 from one partner and helices 3 and 4 of the opposite partner. The client protein binding site of 14-3-3 is located within an amphipathic groove formed by helices 3, 5, 7, and 9 (Figure 35.1). Since each individual unit of a 14-3-3 dimer can bind one client protein, 14-3-3s can act as a docking station, bringing together two different clients. Alternatively, a 14-3-3 dimer can attach to two binding sites on the same client protein and thus might change the conformation of that protein. Divalent cations or polycations, such as polyamines, may often enhance the binding of 14-3-3 to client proteins [17]. The 14-3-3 binding motif in client proteins has been analyzed by the use of degenerate peptide libraries [18, 19] and found to consist of short conserved phosphorylated sequences, termed mode I (RSXpSXP) or mode II (RSX ϕ XpSXP) in which pS represents a phosphorylated Ser residue and ϕ is an aromatic or aliphatic amino acid. This sequence is not invariable—for example, Lys may substitute for Arg, as is seen for *Arabidopsis* GAPDH and pThr for pSer (*Arabidopsis* Gln tRNA synthase). Such a small, variable sequence is not infrequently encountered, and indeed it has been estimated that some 30–40% of *Arabidopsis* proteins could potentially function as 14-3-3 clients [5]. Of course, not all of these will interact with 14-3-3 because (i) the site must be sterically available, (ii) the site must be phosphorylated, and (iii) both 14-3-3 and client must be present at the same time and place. A new phosphorylated mode III motif located at the C-terminus of some proteins has recently been described, the consensus being p(S/T)X1-2-COOH [20]. To add to this complexity, a number of nonphosphorylated binding sites have been characterized [for example see reference 21]. Both phosphorylated and nonphosphorylated target sites appear to bind to the same 14-3-3 region. Secondary structure predictive analysis of 14-3-3 binding proteins indicates that the binding site is generally located in a region with high levels of intrinsic disorder [22]. The number of plant client proteins that 14-3-3 has been shown to interact with has recently rapidly expanded, as a direct consequence of improvements in modern proteomic techniques. Specificity of 14-3-3 isoform binding to individual client proteins

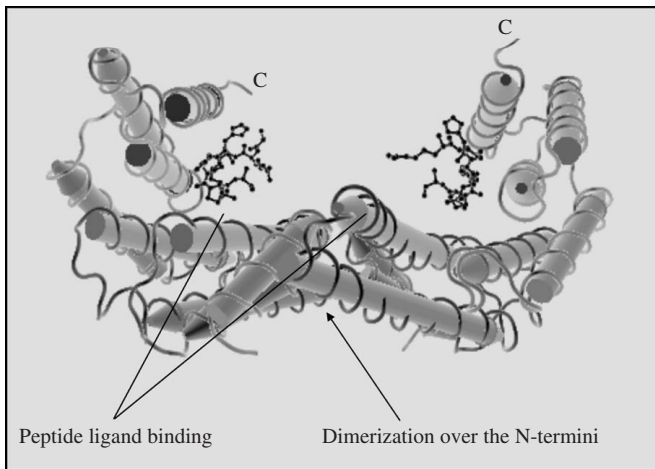


FIGURE 35.1. A 14–3–3 dimer with bound peptide ligands at the binding sites. The two C-termini (C) are indicated, as is dimerization over the N-termini, although it should be noted that the complete full-length C and N termini are not shown. 14–3–3 proteins are shown as helices, and ligands are shown as ball-and-stick models. Created from mmdbld:36278 using Cn3D.

has often been demonstrated, but the basis for this is not easy to determine since the binding site within 14–3–3 is much conserved between isoforms and binding analysis of 14–3–3 isoforms to phosphopeptide libraries showed little discrimination. Binding specificity may depend on a number of factors including post-translational phosphorylation of the client protein or of 14–3–3 [23], C-terminal inhibition of binding [24, 25], specific amino acid sequence variations in loop 8 of the C-terminal [26, 27], tissue-specific expression and subcellular location of both 14–3–3 and client protein [27, 28], and the steric accessibility of binding sites.

14–3–3 Proteins in Plants

Much of what we have learned about 14–3–3 in plants has come from studies of interactions with one of the first plant proteins to be identified as a 14–3–3 client, the enzyme nitrate reductase (NR [reviewed in references 29 and 30]), although other intensively studied interactions, such as with H⁺ ATPase [31] and SPS [32], have also greatly aided our understanding. Nitrate reductase is a key enzyme in plant nitrogen metabolism, converting nitrate to nitrite. In the light, this process is coupled to amino acid synthesis, as reduced ferredoxin is required. In the absence of the photosynthetic process, nitrite would accumulate, resulting in cytotoxicity; thus NR is inactivated in the dark through phosphorylation of a conserved seryl residue and the binding of an NR inhibitor protein (together with divalent cations or polyamines). The NR phosphoprotein-binding inhibitor protein was subsequently identified as 14–3–3 [33, 34]. Several different NR kinases have been characterized, including an SNF1-like kinase and two CDPKs. Interestingly, these protein kinases can also be bound by

14-3-3, but the significance of this is not yet clear. Reactivation of NR activity is by dephosphorylation of the regulatory Ser, but exactly how this occurs is also not understood since bound 14-3-3 physically blocks dephosphorylation; it is possible that 14-3-3 is post-translationally modified to change its binding properties, thus allowing phosphatases to act on the binding site. NR is also proteolytically degraded in the dark, and 14-3-3 binding might be a prerequisite for this. Binding specificity of 14-3-3 to NR was studied in barley and depends on isoform expression patterns in the plant and on the amino acid sequence in loop 8 (between helix 8 and 9) of 14-3-3 [27]. The presence of a specific Gly residue in this cation-binding loop confers additional flexibility to the C-terminal, which can alter target binding and also renders the C-terminal more susceptible to proteolysis [26]. The C terminal of the *Arabidopsis* ω isoform may act as an autoinhibitor of binding, possibly by physically blocking access to the amphipathic groove [35]. Interestingly, the C-terminus of barley isoform A is specifically cleaved *in vivo* during germination [24]. This proteolytic processing may therefore also contribute to binding specificity [25].

35.2 METHODOLOGY AND STRATEGY

Different experimental objectives have been addressed by the methods described here: first, to examine the 14-3-3 isoform proteome in a given species; second, to screen for novel interacting partners; and third, to focus on the individual client protein, in order to investigate the mechanisms of 14-3-3 protein binding. Examples of these strategies are given in the next section. Critical to the approaches involving PPIs is the correct use of controls to show specificity of interaction. Thus, use of unrelated proteins, or proteins mutated in their binding sites, shows specificity of interaction rather than “sticky” proteins. Since the most common 14-3-3 binding sites are phosphorylated, client binding should be phosphorylation-dependent; therefore treatment of client proteins with phosphatase should greatly reduce binding. The *Arabidopsis* 14-3-3 proteome has been analyzed by 2-DGE fractionation of cell extracts using a pH 3–6 IPG strip followed by immunoblotting of the 2D gels with a 14-3-3 specific antibody to identify the family of isoforms. This was followed by endoprotease digestions, HPLC, and ESI-MS/MS of peptides for isoform identification [15]. Similarly, *Arabidopsis* suspension-cultured cells extracts were fractionated by column chromatography followed by 2-DGE and trypsin digestion and MALDI-TOF-MS analysis for identification of 14-3-3s [36]. A quantitative proteomics approach has employed SIL of *Arabidopsis* for an analysis of 14-3-3 binding to the NR and SPS 14-3-3-binding motifs [37]. Another analytical tactic has been to create transgenic plants designed to express specific 14-3-3 genes as GFP fusions in order to monitor the subcellular location of the proteins [38].

The Y2H system [39] has been used both to screen for novel interacting partners and to analyze the binding of a specific client protein. To search for novel partners, an activation domain-tagged cDNA library in yeast is needed, which allows many interactions to be quickly screened. However a careful analysis of positives from the Y2H is required because overexpression of proteins can yield artifactual associations.

Far Westerns or 14–3–3 overlays have been used to visualise 14–3–3 binding on blots. Here, plant extracts are separated by SDS-PAGE, transferred to nitrocellulose membranes, and hybridized with digoxigenin-labeled 14–3–3, which can be detected immunologically. The use of phosphopeptides corresponding to a strongly binding 14–3–3 site (e.g., ARAApSAPA) to outcompete the binding illustrates the specificity of the interaction. Co-immunoprecipitation with 14–3–3 antisera also has been used to monitor protein interactions with 14–3–3 [40].

The use of immobilized recombinant 14–3–3 for affinity purification from plant extracts was first described by reference 40. In these experiments, bound proteins from cauliflower extracts (this was used because bulk quantities of tissue are available from a close relative of *Arabidopsis*) were bound to immobilized yeast 14–3–3 and specifically eluted with a 14–3–3 binding site phosphopeptide. The eluted proteins were subjected to anion exchange chromatography and SDS-PAGE, followed by excision of stained protein bands and trypsin digestion. Peptides were then separated by RP chromatography and sequenced in order to identify the proteins. In a similar approach, the spectrum of barley grain client proteins binding to barley 14–3–3A protein was investigated by immobilizing the recombinant 14–3–3 to sepharose and using this matrix in affinity chromatography experiments. In order to simplify separation and analysis of bound proteins, the grain extracts were fractionated by ion exchange chromatography prior to affinity chromatography. After extensive washing of the affinity matrix, proteins were eluted with SDS-PAGE sample buffer, fractionated by SDS-PAGE, CBB stained, and identified by MALDI-TOF-MS [41]. The presence of a potential 14–3–3 binding site in a given protein sequence can also be used to predict binding of that protein. The limit dextrinase inhibitor protein contains a mode I motif and has been shown to bind 14–3–3 in a phosphorylation-dependent manner by immunoblotting after 14–3–3 affinity chromatography [42].

35.3 EXPERIMENTAL RESULTS AND APPLICATIONS

Due to its relative abundance in all cell types, 14–3–3 has been fortuitously identified in a number of plant proteomic studies not directly aimed at 14–3–3 [for example, see reference 43]. However, the plant 14–3–3 proteome has been specifically investigated in *Arabidopsis* 14–3–3 suspension culture cells. Proteins encoded by 11 of the 13 *Arabidopsis* genes were positively identified. All the N-termini of analyzed isoforms were found to be acetylated (as is the case for mammalian 14–3–3), which may play a role in promoting protein stability. Other modifications such as phosphorylation or the C-terminal cleavage, as described for barley 14–3–3A, were not found, but could not be ruled out in this study [15]. Nine isoforms were also identified by MALDI-TOF-MS from *Arabidopsis* suspension culture cells after partial purification and 2-DGE [36]. Subcellular location of 14–3–3 has been investigated by a number of means including the use of GFP fusions in transgenic plants; this study showed that different isoforms showed distinct subcellular locations, which could be disrupted by means of chemical agents that intervene with 14–3–3/client binding. This confirms that individual isoforms have binding specificity and shows that subcellular location of 14–3–3 is driven by interactions with the binding partners.

The isoform specificity of tobacco 14-3-3 binding to SPS has been analyzed by screening for interacting partners in a Y2H assay; 12 SPS-interacting clones were identified, representing two 14-3-3 isoforms. Deletion analysis of these clones indicated that the C-terminus of 14-3-3 mediates binding specificity [44]. Y2H analysis was also used to investigate the interactions between barley 14-3-3 and ABA-responsive transcription factors of the ABI5 family, clearly illustrating the isoform specificity of these interactions [45].

Affinity purification of cauliflower proteins to immobilized yeast 14-3-3 and amino acid sequencing of peptide fragments resulted in the identification of nine different proteins, including enzymes of carbohydrate metabolism, nitrogen metabolism, and also a protein involved in floral development [40]. A similar approach, together with 14-3-3 overlays, was used to analyze 14-3-3 binding proteins from *Arabidopsis* cell suspension cells, which uncovered a nutrient sensing role for 14-3-3s. Sugar starvation of the suspension culture cells resulted in loss of 14-3-3 binding and the proteolytic cleavage of the client proteins; 14-3-3 binding is thought to protect these proteins from cleavage [46]. Also using an affinity chromatography approach, the binding of barley grain proteins to immobilized barley 14-3-3A was shown to be phosphorylation-dependent since hardly any binding was detected when extracts were dephosphorylated with AP [41]. After extensive washing, 54 of the binding proteins were identified, of which the largest group (31%) was involved in carbohydrate metabolism (in some respects not surprising because this is the major function of the barley grain). Four of the five enzymes involved in sucrose biosynthesis (TPI, FBPase, S(6) PS, and S(6) PS phosphatase) were identified, which suggests the coordinated regulation of sucrose biosynthesis by 14-3-3 proteins. For the enzyme sucrose synthase, it could be demonstrated that activity in barley grain extract was reduced by addition of exogenous 14-3-3. A similar example is shown in Figure 35.2, where the activity of barley grain soluble and cell-wall-bound acid invertase in crude extracts is reduced by addition of recombinant 14-3-3A.

As well as proteins concerned with metabolism, a large functional group of 14-3-3 clients was composed of proteins related to abiotic and biotic plant defense. Of particular interest are proteins of the NBS-LRR class, characteristic of plant disease resistance genes (*R* genes) encoding receptors to pathogen avirulence gene products; three proteins belonging to this category were found. It was previously found that 14-3-3 gene expression is up-regulated by powdery mildew infection [11], and the *Mla* powdery mildew resistance locus is also known to encode an NBS-LRR protein. This suggests an important role for 14-3-3 in the regulation of disease signal transduction for powdery mildew and perhaps for other pathogens. The barley NBS-LRR proteins described in the public databases were screened for potential 14-3-3 binding sites, and at least 30 were found that contained such motifs, sometimes in many copies (up to 20 motifs in one protein).

Affinity purification experiments on 14-3-3 client proteins from plant extracts have isolated predominantly proteins concerned with cellular metabolism. It was once thought that there is a marked difference in the type of client protein in animal and plant cells, with 14-3-3 clients from animal cell being primarily proteins concerned with cell signaling; however, this difference may be attributable to the techniques

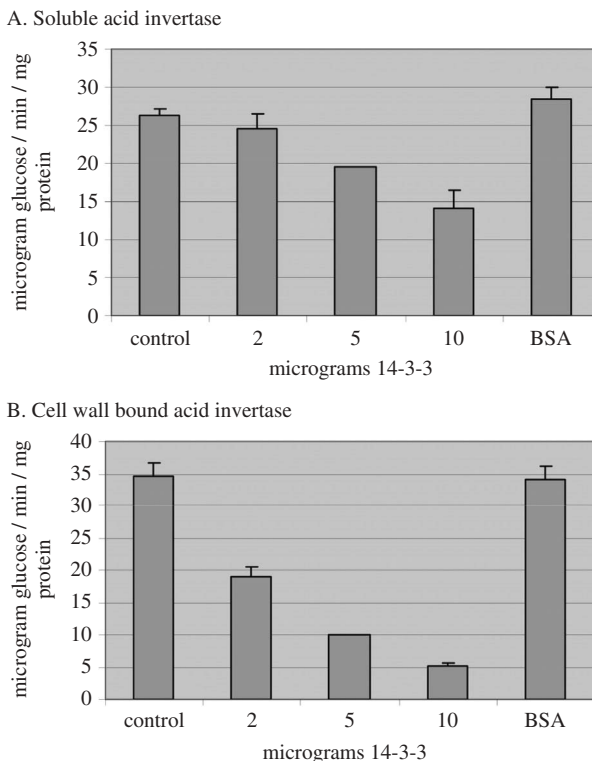


FIGURE 35.2. Effect of barley 14–3–3A on acid invertase activity in barley endosperm extracts. Developing endosperm extracts **A** (soluble, 60 μ g protein) and **B** (cell wall bound, 35 μ g protein) were preincubated for 30 mins with ATP (1 mM) and increasing concentrations of purified recombinant barley 14–3–3A protein or 10 μ g BSA, and then invertase activity (release of glucose from sucrose) was assayed. Control: no 14–3–3 added. Data represent the mean of three replications. Error bars indicating the SD (Ross and Morris, unpublished data).

employed; mammalian 14–3–3 workers have made extensive use of Y2H screens. Thus when affinity purification was used to isolate 200 14–3–3 binding proteins from mammalian cells, the largest group of proteins found (24%) were enzymes of primary and secondary metabolism; only about 15% of the proteins were associated with cell signaling, and 15% with TFs and RNA-binding proteins [47].

This bias of protein functional class being dependent on the analytical methodology was well-illustrated in a recent analysis of barley proteins. The Y2H system was directly compared with affinity chromatography of 14–3–3 binding partners from barley seedling [48]. One hundred thirty-two putative targets were isolated by the Y2H method and 30 by affinity chromatography. About 10% of the partners isolated by the Y2H method were also found by affinity chromatography. An interesting difference in the functional classification of interacting partners obtained from both approaches was noted; affinity purification yielded predominantly (44%) proteins

concerned with metabolism, whereas the Y2H method found a large proportion of cell-signaling-related proteins (35%). One reason for this difference is that affinity purification will yield results that reflect the relative abundance of those proteins in the cell; proteins involved in signaling are much less abundant than those involved in cellular metabolism. In the Y2H method, the cDNA clones are all expressed from the same promoter, so that protein levels in the yeast cells will be approximately equalized; however, there will still be a bias in favor of those clones that are more highly represented in the cDNA library.

Interacting proteins can be analyzed in a more indirect manner by manipulating 14-3-3 expression in transgenic plants and monitoring changes in enzyme activity and metabolites in the transgenics. This approach has been successfully carried out in potato, where expression of specific 14-3-3 isoforms was down-regulated by antisense [49]. The most significant results found here were enhancements in NR and SPS enzyme activity, thus confirming previous *in vitro* data on these interactions.

35.4 CONCLUSIONS

Our understanding of the 14-3-3 family and the 14-3-3 client proteome in plants has made great progress in recent years, due to the availability of robust analytical methods for protein fractionation and characterization and, equally importantly, easy access to an increasingly comprehensive sequence database, which permits protein identification. Some 200 individual plant 14-3-3 client proteins are now described in the literature, although for the majority of these, characterization is still limited to evidence of 14-3-3 binding, either *in vitro* or in a Y2H system. These client proteins include many centrally important enzymes of carbon and nitrogen metabolism. Entire metabolic pathways are represented, for example, nearly all the enzymes for sucrose biosynthesis [41], and glycolysis [48]. This suggests that these pathways might be integrated into multicellular complexes, with 14-3-3 as a coordinating scaffold or regulatory protein. Evidence is also accumulating for 14-3-3 proteins playing a role in plant defense mechanisms and in plant hormone action. However, although we have a good picture of some of the 14-3-3/client interactions (for example NR, SPS), important questions remain for most of the others; how big is the 14-3-3 client proteome, what are the dynamics of the binding proteome during development and disease, what are the biological consequences of 14-3-3 binding, how is it regulated, and how is specificity of binding ensured? Importantly, how can we put these findings to good use in improving plant health and productivity?

35.5 FIVE-YEAR VIEWPOINT

We are still a long way from an understanding the full complement of 14-3-3 interacting proteins in plants. Clearly, in the next five years, improvements in proteomics technology, coupled with more completed plant genomes and better annotation, will aid this effort, as is true for other plant proteomics projects. Specifically for

PPI analysis, techniques such as TAP-tagging, expression of epitope-tagged 14–3–3 followed by co-immunoprecipitation, and BN-PAGE will also assist in the analysis of multiprotein complexes. But this is only one side of the story. The questions of 14–3–3 isoform binding specificity will also need to be addressed. 14–3–3 generally binds to a phosphorylated motif; the protein kinases that carry out this phosphorylation, as well as the protein phosphatases that reverse it, are critically important to the process, as is the regulation of their activity. These will be prime targets for research in the near future. Since 14–3–3 binding is known to be an integral part of carbon and nitrogen metabolism, and there is increasing evidence for a role in plant hormone action and in defense responses, a further challenge is to use this information in order to steer plant metabolism in the direction of improved yield and quality of crop plants.

REFERENCES

1. Aitken A. 2006. *Semin Cancer Biol* **16**:162–172.
2. Ferl RJ. 2004. *Physiol Plant* **120**:173–178.
3. Ferl RJ, Manak MS, and Reyes MF. 2002. *Genome Biol* **3**:reviews 3010.1–3010.7.
4. Roberts MR. 2003. *Trends Plant Sci* **8**:218–223.
5. Sehnke PC, DeLille JM, and Ferl RJ. 2002. *Plant Cell* **14** Suppl: S339–S354.
6. Moore B and Perez V. 1967. Specific acidic proteins of the nervous system. In: *Physiological and Biochemical Aspects of Nervous Integration* (Carlson FD, ed.), Prentice-Hall, Englewood Cliff, NJ, pp. 343–359.
7. Ichimura T, et al. 1987. *FEBS Lett* **219**:79–82.
8. Hirsch S, et al. 1992. *FEBS Lett* **296**:222–224.
9. de Vetten NC, Lu G, and Ferl RJ. 1992. *Plant Cell* **4**:1295–1307.
10. Lu G, et al. 1992. *Proc Natl Acad Sci USA* **89**:11490–11494.
11. Brandt J, et al. 1992. *Plant J* **2**:815–820.
12. van Heusden GP and Steensma HY. 2006. *Yeast* **23**:159–171.
13. Fu H, Subramanian RR, and Masters SC. 2000. *Annu Rev Pharmacol Toxicol* **40**:617–647.
14. Chen F, Li Q, Sun L, and He Z. 2006. *DNA Res* **13**:53–63.
15. Fuller B, et al. 2006. *Proteomics* **6**:3050–3059.
16. Xiao B, et al. 1995. *Nature* **376**:188–191.
17. Athwal GS and Huber SC. 2002. *Plant J* **29**:119–129.
18. Muslin AJ, et al. 1996. *Cell* **84**:889–897.
19. Yaffe MB, et al. 1997. *Cell* **91**:961–971.
20. Coblitz B, et al. 2005. *J Biol Chem* **280**:36263–36272.
21. Petosa C, et al. 1998. *J Biol Chem* **273**:16305–16310.
22. Bustos DM and Iglesias AA. 2006. *Proteins* **63**:35–42.
23. Dubois T, et al. 1997. *J Protein Chem* **16**:513–522.
24. Testerink C, et al. 2002. *Plant Mol Biol* **50**:535–542.
25. Truong AB, et al. 2002. *Proteins* **49**:321–325.
26. Sehnke PC, et al. 2006. *Plant Physiol* **140**:647–660.

27. Sinnige MP, et al. 2005. *Plant J* **44**:1001–1009.
28. Testerink C, et al. 1999. *Plant Physiol* **121**:81–88.
29. Lillo C, et al. 2004. *J Exp Bot* **55**:1275–1282.
30. MacKintosh C and Meek SE. 2001. *Cell Mol Life Sci* **58**:205–214.
31. Bunney TD, van Walraven HS, and de Boer AH. 2001. *Proc Natl Acad Sci USA* **98**:4249–4254.
32. Toroser D, Athwal GS, and Huber SC. 1998. *FEBS Lett* **435**:110–114.
33. Bachmann M, et al. 1996. *FEBS Lett* **387**:127–131.
34. Moorhead G, et al. 1996. *Curr Biol* **6**:1104–1113.
35. Shen W, Clark AC, and Huber SC. 2003. *Plant J* **34**:473–484.
36. Harthill JE, et al. 2006. *Plant J* **47**:211–223.
37. Gruhler A, et al. 2005. *Mol Cell Proteomics* **4**:1697–1709.
38. Paul AL, Sehnke PC, and Ferl RJ. 2005. *Mol Biol Cell* **16**:1735–1743.
39. Fields S and Song O. 1989. *Nature* **340**:245–246.
40. Moorhead G, et al. 1999. *Plant J* **18**:1–12.
41. Alexander RD and Morris PC. 2006. *Proteomics* **6**:1886–1896.
42. Stahl Y, et al. 2007. *Plant Sci* **172**:452–461.
43. Schiltz S, et al. 2004. *Plant Physiol* **135**:2241–2260.
44. Börnke F. 2005. *J Plant Physiol* **162**:161–168.
45. Schoonheim PJ, et al. 2006. *Plant J* **49**:289–301.
46. Cotellet V, et al. 2000. *EMBO J* **19**:2869–2876.
47. Pozuelo Rubio M, et al. 2004. *Biochem J* **379**:395–408.
48. Schoonheim PJ, et al. 2006. *Plant Physiol* **143**:670–683.
49. Zuk M, Weber R, and Szopa J. 2005. *J Agric Food Chem* **53**:3454–3460.

PART VI

MULTIPROTEIN COMPLEX

TAP-TAGGING SYSTEM IN RICE FOR PROTEIN COMPLEX ISOLATION

Jai S. Rohila and Michael E. Fromm

36.1 INTRODUCTION

Rice feeds more than one-third of the global population and is an important model cereal crop due to its 389-Mb genome [1] and convenient gene transfer methods [2]. About 45,000 protein-coding sequences have been predicted in rice haploid genome [3]. Studying the functions of rice genes is an active research area using primarily genetic methods. An additional area of research is discovering protein interactions, because they can reveal some insights into a protein's function and *in vivo* mechanisms [4]. A recently developed TAP-tag technology is a state-of-the-art methodological approach [5] for large-scale analyses of protein complexes *in vivo* [6, 7] and is comparatively easy for researchers with a limited experience in protein purifications. TAP-tag technology has been successfully used in yeast [6], *Drosophila* [8], mammalian cells [9], *E. coli* [10], *Chlamydomonas* [11], and higher plants [12–15]. In TAP the protein complex is sequentially affinity purified, and the proteins present in the complex are identified with the help of MS and bioinformatics.

The TAP-tag was first used in yeast [16] for large-scale PPI studies. The first successful report on the use of TAP-tag technology for the purification of a protein complex *in planta* was demonstrated by our group [12]. The TAP-tag can be

fused on either C- or N-terminus to the protein of interest and are available in vectors incorporating GATEWAY™ (Invitrogen) recombination sites for expression of fusion proteins under the control of the cauliflower mosaic virus (CaMV) 35S promoter. The usefulness of the TAP-tag is further demonstrated by analyzing over 100 TAP-purified-protein kinase purifications [13, 15] in our lab.

There are several other versions of TAP tags available. Among these, ProA-3C protease-6xHis:9xmyc [14], BirA (biotin):target protein:TAP-tag [17], ProA-TEV-FLAG [9], and CBP:6xHis:3xHA [18] are available from academic institutions, while InterPlay™ TAP System (SBP:CBP) (Stratagene) and FLAG®HA TAP kit (Sigma) are distributed by companies. In a complementation study of mutant lines with seven TAP-tagged proteins, it was shown that three did not complement, two partially complemented, and two fully complemented the mutant phenotypes [14]. This indicates that the expression and function of TAP-tagged proteins are specific to each fused protein. The probability of success with the method will be discussed more fully below.

36.2 METHODOLOGY AND STRATEGY

TAP Methodology Overview

A TAP project starts with making a gene construct that fuses the TAP-tag to a target protein of interest. This construct is then transformed into the plant, and transgenic plants expressing the fusion protein are identified. Transient expression is readily detectable in *N. benthamiana* agroinfection assays [19] and can be useful as a quick check that the protein fusion is expressed and functional. Crude protein extracts from the initial transgenic plant, or more typically the progeny plants, are prepared from these transgenic cells. The fusion protein as well as interacting proteins present in a protein complex are then recovered by TAP. The TAP method involves two sequential affinity purification steps: an IgG column followed by CaM column purification. The two purification steps are followed by separation of the individual purified proteins by SDS-PAGE. The protein bands are visualized with a fluorescent dye, excised, and analyzed by MS to identify the proteins. The TAP strategy has been summarized in Figure 36.1.

TAP-Tag Structure

Typically, a TAP-tag consists of two protein-binding domains separated by a protease cleavage site. The Protein A domain from *S. aureus* (ProA) and CaM binding protein (CBP) domain in the TAP-tag have sufficient affinity for efficient recovery of a fusion protein present at a low level in a complex crude extract from a transgenic plant (<http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/ExternalInfo/seraphin/TAP.html>). These two affinity tags, ProA and CBP, are separated by a TEV protease cleavage site. TEV cleavage is used to separate the ProA domain, which remains tightly bound to the IgG beads, from the released protein complex. The recombinant TEV protease recognizes the ENLYFQG protein

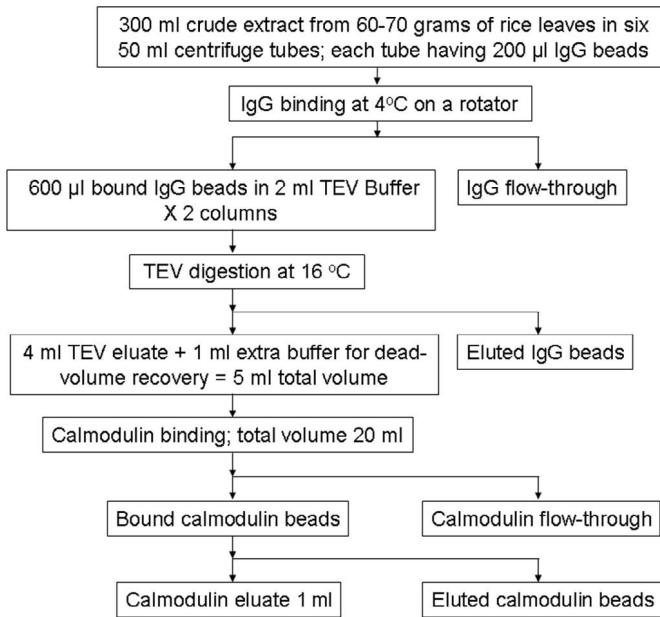


FIGURE 36.1. An overview of the steps in TAP.

sequence in TAP-tag, and the cleavage occurs between Q and G amino acid residues (Figure 36.2).

The N-terminal and C-terminal TAP-tags have reverse orders of the domains in the TAP-tag to allow for the TEV cleavage of only the ProA domains. These synthetic TAP-tags also include the castor bean catalase intron1, which is known to increase gene expression in rice plants [20]. The TAP-tag cassettes were cloned into a binary vector containing gateway recombination sites (Invitrogen) either N-terminal to the attR1 sites or C-terminal to the attR2 site. The N-terminal TAP-tag is designed to be in reading frame 1 of the downstream attB1 site, and thus in this case the AUG start codon of the target gene is separated from the 3' end of CBP domain by a 19-amino-acid structure consisting of exons flanking the catalase intron, junction regions, and attB1 site. The C-TAP tag is also designed to be in reading frame 1 upstream of the attB2 site, and so in this case a 20-amino-acid structure consisting of attB2, junction region, and catalase exons separates the 3' end of the target gene and the 5' end of the CBP domain (Figure 36.3). These 19- or 20-amino-acid linkers serves as hydrophilic spacers between the TAP-tag and the target protein, presumably increasing the steric availability of the CBP region for binding to CaM beads. The N-TAP and C-TAP tag-Gateway cassettes were cloned in the pPTN289 (Tom Clemente, unpublished data) binary vector. The N-terminal or C-terminal TAP-tag fusion cassettes expressed from ECaMV 35S promoter are available as HindIII restriction fragments for transfer to other vectors. In our lab this TAP-tag fused with different rice protein kinases has been expressed in rice using the maize Ub promoter.



FIGURE 36.2. Structure of N-terminal and C-terminal TAP-tags.

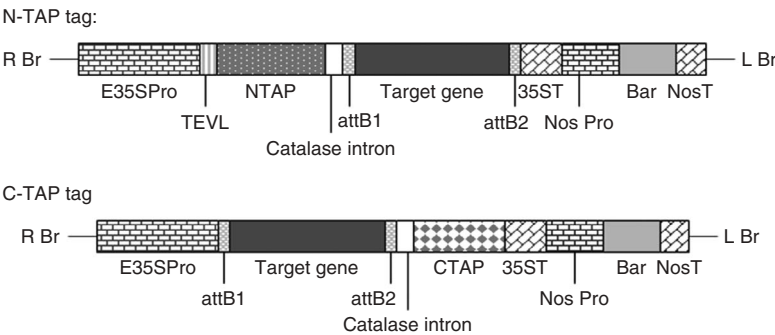


FIGURE 36.3. Structure of the T-DNA expressing TAP-tagged target protein. The ECaMV 35S promoter (E35SPro) and CaMV 35S polyadenylation region (35ST) are used to express the TAP-tagged gene. TEVL in N-TAP tag figure designates the TEV untranslated leader. R Br and L Br represent right and left borders, respectively. See insert for color representation of this figure.

Production, Checking of Transgenic Plant Material, and Protein Purification Procedure

Transient and Stable Expression Approach. In our lab we have isolated protein complexes from both transient [12] as well as stably expressed TAP-tagged proteins [13]. Transient expression approach is quick and is useful for confirming

expression of the fusion protein if the promoter is expressed in the transient system. Stably transformed plants are essential for analyzing the protein complex under conditions more similar to those of in the native plant.

We initially used transient assay systems to verify the TAP-tag system was useful in plants. Using a transient approach, we expressed a TAP-tagged synthetic glucocorticoid receptor (GVG) in *N. benthamiana* leaves. GVG protein is located in the cytoplasm until released by the glucocorticoid dexamethasone [21]. Once released, it enters the nucleus to activate transcription of its target promoter. In mammalian cells, the glucocorticoid receptor is bound to HSP90 in the cytoplasm. Our TAP-purified GVG protein was found to be interacting with HSP70 and HSP90 in plants [12]. Witte et al. [22] also report a successful purification of TAP-tagged AtSGT1b from *N. benthamiana* leaves using transient approach. We have also purified over 100 TAP-tagged protein kinases from stably transformed rice plants with the success rates shown in Table 36.1 [13, 15].

Checking the Expression of the Fused Protein

CaM-Bead Assay. The CBP domain of the TAP-tag provides a convenient bead-based assay for expression levels. The protocol is essentially same as described by Rohila et al. [12]. Briefly, protein extracts are prepared in CaM-binding buffer (CBDB) and bound to CaM-agarose beads for 1 h. The bound-CaM beads were washed in CBDB and incubated in CBDB supplemented with 30 mg/mL gelatin and a 1:5000 dilution of PAP conjugate (Sigma) for 1 h at room temperature. The POX protein in the PAP conjugate is similar to a conjugated secondary antibody and can be detected with the same ECL reagents. The beads were washed thrice in CBDB and once with the substrate prior to adding the substrate (Supersignal® west pico chemiluminescent substrate; Pierce). The signal was detected in a luminometer (Lumistar; BMG Labtechnologies). For consistency in the results, always include a positive and a negative control in the experiment. Here the intensity of luminescence signal positively correlates with the expression of the TAP-tagged protein in the plant tissue.

Western Blot. For immunodetection of TAP-tagged protein, either the ProA domain or CBP domains can be used. Detection of the ProA domains is much more sensitive and easier using the PAP conjugate to detect the ProA domain of the TAP-tag as described by Rivas et al. [23]. It should be noted that the ProA domain cannot

TABLE 36.1. Summary of 129 Tandem Affinity Purifications

	Number	Percent
Recovered TAP-tagged protein kinases	94/129	73
Identified complexes ^a	67/94	71
Complexes with high degree of confidence	10/94	11

^aExcluding contaminant/reoccurring proteins

be used for immunodetection once the fused protein is digested with TEV. After TEV digestion the researcher can detect the CBP domain of the TAP-tag on the Western blots, if the tagged-protein specific antibodies are not available. Additionally, if using the C-terminal TAP tag, an antibody is commercially available to detect the SSGALDYDIPTTASENLYFQ sequences of the C-TAP-tag (Open Biosystems).

In our lab the CBP domain was detected by Western blots using biotinylated-CaM [24]. For this the electroblotted membrane was blocked overnight in Tris-buffered saline (TBS) containing 1% BSA at 4°C. The membrane was washed in Buffer A followed by Buffer B and probed with 450 ng/mL of biotinylated CaM (Calbiochem) in Buffer B containing 1% BSA for at least 2–3 h. The membrane was washed in Buffer A and incubated in a 1:3000 dilution of streptavidin-AP (Amersham Biosciences) in Buffer B containing 1% BSA for 1 h at room temperature. After washing in Buffer B, the membrane was developed using 1-Step™ NBT/BCIP (Pierce).

Buffer A

20 mM Tris-HCl, pH 7.5 (2.5 mL of 1 M stock)

150 mM NaCl (15 mL of 1 M stock)

1 mM CaCl₂ (100 µL of 1 M stock)

50 mM MgCl₂ (5 mL of 1 M stock)

0.01% Tween 20 (100 µL)

H₂O to 100 mL final volume

Buffer B

Same as Buffer A but without Tween 20

Preparation of Crude Extract for TAP

Preparation of extract is a critical step in the purification process, since this step influences the total quantity of desired protein in the crude extract and the complex recovered and the biological activity of the protein. A number of variables determine the success of a protein extraction procedure: grinding of the plant tissue, presence of protease inhibitors, choice of buffers, working temperature, and so on. We grind frozen rice leaves in a high-speed metal blender until they produce a fine powder in liquid nitrogen. We routinely use the buffer composition mentioned in this chapter for rice tissue to prepare the crude extracts after grinding. As a thumb rule, we use at least 4 mL of PEB per 1 g (approximately) of rice tissue for TAP. The crude protein extract is first cleared by passing it through a fine sieve and then passed through two layers of miracloth which is followed by two centrifugations at 4°C at 17,000 rpm for 25 min each. The cleared protein extract is ready for the IgG binding step of the purification. Less stable protein complexes may require more rapid purification steps. One centrifugation is adequate in this case.

Purification of the Fused Target Protein and Associated Complex

Initially, optimal conditions for every step of the purification were determined in pilot-test experiments. These optimized conditions were combined to generate the final optimized TAP protocol that is being described here. We have purified more than 100 different TAP-tagged protein kinases in rice with about a 70% success rate of recovering the tagged kinase using this TAP protocol [13, 15]. The failures were due to either very low expression of the fused protein, as measured using the bead assay or Western blots, or *in vivo* proteolysis of the TAP-tag.

The protein extract prepared as above is incubated with washed and equilibrated IgG beads (Amersham Biosciences) on a rotary shaker at 4°C for 2 h. Then the bound IgG beads were washed extensively first with PEB and then briefly with TCB. Bound IgG beads are incubated with 150–200 units of rTEV protease (Invitrogen) for 1–1.5 h at 16°C to specifically release the bound material from IgG beads. The IgG-eluate is incubated with equilibrated CaM beads in the presence of CBDB for 1 h at 4°C. This second affinity step removes the TEV protease as well as traces of other contaminants remaining after the first affinity purification. The bound CaM beads are washed extensively in CBDB, and then the proteins are specifically released in four elutions using 250 µL of EGTA buffer each time. These fractions are pooled and TCA precipitated by adding 335 µL of 6.1 N TCA solution (Sigma) in 1 mL of CaM eluate and incubated for 1–2 h on ice. After precipitation, the mixture was centrifuged at 14,000 rpm for 30 min at 4°C. The resulting pellet was washed in freezing-cold (bottles kept at –20°C freezer) 1 mL of acetone + 0.05 N HCl followed by second wash in freezing-cold acetone only. The pellet was air-dried for about 2–3 min on bench. The precipitated protein was dissolved in 32 µL of 20 mM Tris-HCl pH 8.0 and mixed with 8 µL of 5X SDS-loading buffer. The mix was boiled for 3–4 min and was loaded onto the SDS-PAGE gel. The protein bands were stained by SRPGS (Bio-Rad) and visualized on a Gel-doc system (Bio-Rad) following manufacture's instructions. SRPGS is fully compatible with MS and allows a sensitive fluorescence detection of proteins in SDS-PAGE [25]. This whole TAP strategy has been described in Figure 36.1.

The following four different kinds of buffers are used in the TAP process. Their recipe is as below:

Protein Extraction Buffer (PEB)

- 20 mM Tris-HCl pH 8.0 (20 mL of 1 M stock)
- 150 mM NaCl (30 mL of 5 M stock)
- 0.1% IGEPAL (1 mL of 10% stock)
- 2.5 mM EDTA (5 mL of 0.5 M stock)
- 10 mM 2-ME (697 µL)
- 20 mM NaF (20 mL of 1 M stock)
- H₂O to 1000 mL final volume
- 2 mM benzamidine, 1 mM PMSF, 1% Sigma's plant protease cocktail, 10 µM leupeptin and 10 µM 3,4 dichloroisocoumarin are added just before use.

TEV cleavage buffer (TCB)

10 mM Tris-HCl pH 8.0 (250 μ L of 1 M stock)
150 mM NaCl (750 μ L of 5 M stock)
0.1% IGEPAL (250 μ L of 10% stock)
0.5 mM EDTA (25 μ L of 0.5 M stock)
H₂O to 25 mL final
1 mM DTT (10 μ L of 0.1 M stock per mL buffer)
1 μ M E-64

Note: Add DTT and E-64 just before use.

Calmodulin-Binding Buffer (CBDB)

10 mM Tris-HCl pH 8.0 (1 mL of 1 M stock)
150 mM NaCl (3 mL of 5 M stock)
1 mM Mg-acetate (100 μ L of 1 M stock)
1 mM imidazole (100 μ L of 1 M stock)
2 mM CaCl₂ (200 μ L of 1 M stock)
0.1% IGEPAL (1 mL of 10% stock)
10 mM 2-ME (70 μ L of stock)
H₂O to 100 mL final volume

CaM Elution Buffer (CEB)

10 mM Tris-HCl pH 8.0 (250 μ L of 1 M stock)
150 mM NaCl (750 μ L of 5 M stock)
1 mM Mg-acetate (25 μ L of 1 M stock)
1 mM imidazole (25 μ L of 1 M stock)
2 mM EGTA (100 μ L of 0.5 M stock)
0.1% IGEPAL (250 μ L of 10% stock)
10 mM β -mercaptoethanol (17.5 μ L of stock)
H₂O to 25 mL final

IgG and CaM Beads Preparation. Use 200 μ L of IgG beads (IgG sepharose Pharmacia) per 50 mL of crude extract for the first step purification and 200 μ L of CaM beads (Stratagene) for the second step purification in TAP. Equilibrate the beads with 1 mL of PEB (for IgG beads) or 1 mL of CBDB (for CaM beads) prior to use.

Binding to IgG Beads. To 50 mL of crude protein extract, corresponding approximately to 10–12 g of leaf tissue, add 200 μ L of IgG beads. In our purifications, we generally use 60–70 g of leaf tissue, which yields about 300 mL of crude extract. We divide this into six tubes of 50 mL each and add 200 μ L of IgG beads in each tube. Rotate for 2 h on a rotisserie shaker (Labquake®) at 4°C.

TEV Protease Cleavage. After 2 h of IgG-binding step, we wash the IgG beads thoroughly in the 50-mL tube itself. After washing, we pool all the IgG beads into two polyprep (Bio-Rad) columns and give another wash with 10 mL of PEB and then with 10 mL of TCB. Now close the bottom plug of the column and add 2 mL of TCB and 200 units of rTEV (Invitrogen) and close the top plug of the column. Rotate both the columns on rotisserie shaker at 16°C for 1.5 h. Now first remove the top plug of the column and then the bottom plug and allow the column to drain by gravity flow into a new 50 mL tube. The dead volume may be eluted with an additional 500 μ L of TEV cleavage buffer. At this stage we pool the IgG eluates from both the columns into one tube, thus having a total 5 mL of TEV eluate for second affinity purification step.

Binding To, and Elution From, CaM Beads. To the 5 mL IgG eluate add 3 volumes of CBDB (15 mL) and 15 μ L of 1 M CaCl_2 to titrate the EDTA coming from the TEV cleavage buffer. Add 100 μ L of washed CaM beads into the solution. After closing the tube, rotate it for 1 h at 4°C on a rotisserie shaker. Now wash the bound CaM beads with 30 mL of CBB and elute 4 fractions of 250 μ L with CEB.

Gel Electrophoresis. The CaM eluates can be either directly subjected to MS [26] or separated on a PAG first. An advantage of SDS-PAGE is the better sensitivity due to less interference from other proteins. TCA precipitation provides an efficient and rapid way to concentrate protein eluates. Dissolve the precipitated proteins in 20 mM Tris-HCl pH 8.0 and load onto a 4–15% gradient SDS-PAG. We use gradient gels as they give higher resolution than with uniform gel concentrations. Proteins are visualized by SRPGS.

MS. MS technique is an essential component of the TAP-tag technology. SDS-PAGE fractionated proteins are first “in-gel” digested with a protease, trypsin is the most commonly used enzyme, but other choices are available [27]. The proteolytic digestion of the proteins yield peptides that can be separated by RP-LC before they go into the mass spectrometer [12, 28]. The resulting MS data can be analyzed with automated protein database search algorithms (e.g., MASCOT). Detected peptides are characterized and mapped to the primary sequence of the protein in a selected database. Data are then searched against the NCBI nonredundant database. If coverage of the map is low due to peptides that are lost during preparation or that are of a size that is not suitable for detection by MS, a parallel proteolytic digestion using a different protease is suggested to produce complementary sequence coverage.

36.3 EXPERIMENTAL RESULTS AND APPLICATIONS

A Case Study of Rice Kinase Complexes

During our protein kinase interaction screening project in rice, we constructed rice protein kinase gene fusions with the TAP-tag, stably expressed them in rice plants, and purified them to identify which other rice proteins are associated with each protein kinase *in planta* (Figure 36.4). The gene fusions of the TAP-tagged proteins in rice

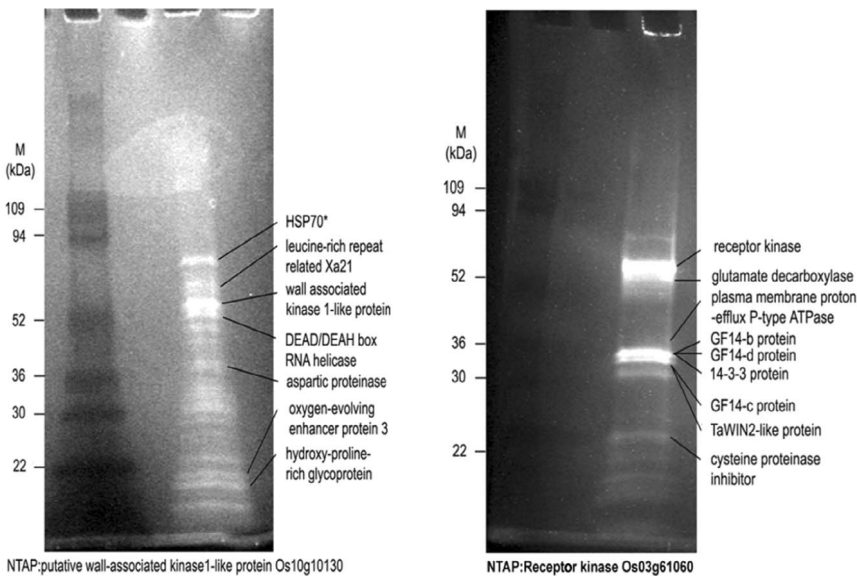


FIGURE 36.4. Example of SYPRO Ruby-stained TAP-purified protein complexes in rice. The left panel shows an ambiguous protein complex, while the right panel is almost stoichiometric and valid.

were expressed from the maize Ub promoter [13]. Shoots of 6- to 8-week-old T-1 transgenic rice plants were harvested and subjected to TAP purifications. TAP data reveals that 73% of the TAP-tagged rice protein kinases can be recovered and 71% of these were isolated as other proteins [13, 15] that are either interacting proteins or co-purifying contaminant proteins as discussed below. A summary of the purification efficiencies is given in Table 36.1.

Examples of some of the interesting protein complexes characterized during TAP in rice were the CDK rCDC2-1 and R2 complex, CK II complex, WAK complexes [13], SNF1-related kinase complex, RLCKVII complex, rice PK1A, PK1B-related protein kinase complex, and a rice CDPK complex [15]. The validity of some of the interacting proteins in the purified complexes was supported by evidences from interactions of similar proteins in other organisms or Y2H experiments. For example, TAP-tagged rice rCDC2-1 has been found to interact with the proteins that are homologous to those of known protein complexes in yeast and human [13]. Interaction data of TAP-tagged casein kinase-II and SNF1 protein complexes were comparable to the interaction data in Y2H screens for the same tagged proteins [29].

Modifications/Alterations to TAP-tag Strategy

Use of Mutants or RNA Silencing with the TAP-Tag. The introduced TAP-tagged fusion protein and the corresponding endogenous protein compete to

participate in protein complexes *in vivo*. TAP-tagged proteins should form protein complexes more efficiently if competition with the endogenous protein can be reduced through the use of RNAi [8] or mutants. Forler et al. [8] used the combination of RNAi and TAP-tag in *Drosophila* cells and called it iTAP strategy. Silencing the endogenous gene without affecting the introduced gene might be difficult when the coding regions are identical. In cases where RNAi knockdowns or mutants are not available, overexpression of the fusion protein can be useful as overexpressed TAP-tagged protein can compete with the endogenous nontagged protein [7, 13].

Adding a Cross-Linking Step Before Purification. A key difficulty with most PPI systems that use crude extracts is determining whether the interaction really occurred in the intact living cells. The crude extracts provide opportunities for proteins to interact that normally might be prevented from interacting in the intact plant cell. A rapid, *in vivo* cross-linking of proteins might provide a potential solution to this problem. An advantage of using cross-linkers could be in giving an extra strength to the weak complexes, ensuring that they do not fall apart during the lengthy purification steps and increases the success rate of purification. A disadvantage is that recoveries are less and the epitopes used in the affinity purification can be modified and not recognized. Several kinds of cross-linkers are available. The most commonly used protein–protein or protein–gene cross-linking reagent is formaldehyde, because it has a short cross-linking distance and is easy to reverse completely by heating the sample at 100°C for 20 min. We have shown previously that 1% formaldehyde cross-linking *in vivo* preserved PPIs between GVG and HSP90 while purifying GVG complex in *N. benthamiana* leaves [12]. However, there are not many reports on combining TAP-tag strategy with cross-linking reagents; it might be a useful strategy when the protein complex is abundant but fragile due to weak interactions. Cross-linking reduces the amount of protein recovered and is difficult to apply to low abundance proteins.

Single Step Affinity Purification. The key issue for different affinity tags is the trade-off between speed and purity. Many protein complexes are not stable enough for the lengthy TAP-tag purification protocol. Faster one-step methods could benefit from an approach, allowing the identification of interacting proteins in a less purified sample using isotope labeling or replicated comparisons to mock purifications. In the case of labeling, the ratio of the labeled proteins is measured using quantitative MS to analyze the composition of a partially purified protein complex and a relevant control purification in which the complex of interest is not enriched. For example, IgG beads can be used to purify proteins from an extract containing a TAP-tagged protein and from an extract containing the wild-type (untagged) version of the protein. Proteins in the two samples are differentially labeled with stable isotopes and analyzed by MS [30]. In this approach, specifically interacting proteins can be distinguished from nonspecifically co-purifying proteins by their abundance ratios. In addition, because complexes can be analyzed after single-step purification, the potential to detect weakly associated proteins is enhanced.

Advantages and Major Concerns of TAP-Tag Strategy

Advantages. The TAP-tag technology is the most exhaustively tested (with over 100 rice protein complex purifications) method available to the plant community which can purify the protein complexes under native conditions and thus ensures that the proteins or even the purified complex can be used for functional assays. However, recovery of high-quality complexes, defined as near-stoichiometric recovery of interacting protein on SYPRO Ruby stained gels, is 11% (Table 36.1). That is, only one of nine tagged proteins is likely to give a high quality result. This seems to be a function of the abundance and stability of the protein complex. This is reflected in the genetic study mentioned above where only two of the seven TAP-tagged proteins could fully complement their corresponding mutants [14]. A reasonable hypothesis is that the basis for the low percentage of good complexes and of genetic complementation is that fusion proteins are often functionally different from the native protein. Alternatively, if this low frequency of high-quality complexes is due to the TAP purification method, there are insufficient studies in plants with other methods for comparison to better define the properties that affect this success rate.

In the cases where a good protein complex is recovered, TAP provides the high degree of purification necessary when a sensitive technique like MS is involved for the identification of proteins in a purified complex. Unlike the Y2H method, the TAP strategy identifies all members of a protein complex in a single purification experiment. Moreover, the TAP system provides an indication of the approximate stoichiometry of the proteins present in a given complex and allows for direct biochemical analysis of the purified proteins [16]. PTMs of proteins, such as phosphorylation, glycosylation, and palmitoylation, plays a critical role in cellular functions; for example, phosphorylation by protein kinases plays a key role in signal transduction pathways in plant cells. The TAP method can also determine PTMs. In our earlier report we provided examples of phosphorylation modifications of the R2, CDC2, and several other rice kinase proteins [13].

Our earlier study [12] shows that TAP typically recovers about 20% of the initial protein, although this is likely to be protein-dependent. The lengthy TAP procedure successfully removes all abundant contaminants to very low levels. Contaminants are variably present at levels of no more than a few nanograms when starting from crude extracts containing 150 mg of soluble protein even when starting with green leaf tissue high in photosynthetic proteins. Recurring contaminants are at very low levels and consist mainly of abundant cellular proteins such as various ribosomal proteins, RuBisCO, GAPDH, glutamate decarboxylase, and PGK [13, 15]. These are readily distinguished from the relatively abundant levels of interacting proteins in a good complex. However, distinguishing minor amounts of interacting proteins from minor amounts of contaminants is difficult and uncertain. One criterion is how often the contaminant appears in a large number of independent purifications as discussed below.

Major Concerns. Any tag, large or small, may influence the property of the tagged protein, which may cause changes in the stability or composition of the protein or protein complex. The TAP-tag size is about 20 kDa and is one of the larger affinity tags. Larger tags are more likely to have steric interference with other proteins in the

protein complex. This effect is likely to be position-dependent, so it is recommended to try both N- and C-terminal TAP-tag fusions in parallel.

One of the key issues in many interaction screens is the difficulty of distinguishing valid interactions from artifactual interactions. In the case of the TAP-tag, artifacts are due to copurification of proteins that do not specifically interact with the TAP-tagged protein. Many of the proteins recovered are at substoichiometric levels and barely detectable by SRPGS. These levels of protein typically give only one or two peptides recovered and identified by MS analysis. High-quality MASCOT scores are sufficient to identify a protein from a single peptide, but the biological significance is ambiguous. That is, the recovered peptide could be from either a valid interaction or a co-purifying contaminant. We have had the advantage of many independent purifications of different TAP-tagged proteins as well as mock purifications to establish a database of commonly occurring proteins recovered that we believe are contaminants. We have defined contaminants as those proteins that are recovered in more than one purifications and are from abundant cellular proteins. It is worth noting that even these abundant cellular proteins are present in the final TAP purifications at almost undetectable levels, as measured by SRPGS. However, the high sensitivity of the MS analysis can still detect these.

The cellular abundance of a protein and its frequency of occurrence in multiple independent purifications can be used to judge which interactions are real. For example, GAPDH, an abundant cellular protein, was found more often in the purifications than annexin protein, and thus GAPDH was put under “contaminant proteins” category. Proteins found in preparations of related tagged proteins are a special case where multiple recoveries of the same interacting proteins are expected. Common interacting proteins also can be recovered in multiple purifications; for example, 14–3–3 protein subunits were found to interact with several phosphorylated TAP-tagged protein kinases [13]. Linker histone H1 and H5 family protein and histone H2A were found together in protein complexes of TAP-tagged protein kinase APK1A and TAP-tagged protein kinase APK1B [15]. When taken altogether, results of this work illustrate the ability of the TAP-tag technology to yield reliable interactors along with novel potential interactors that need further verifications and characterization.

Ambiguous cases require additional confirmation, but the choice of how to do this is not simple. The TAP-tag purification is essentially a high-sensitivity “pulldown” experiment. That is, one protein is tagged and the other is detected by purification and MS analysis instead of by Western blot as is typical in pulldown experiments. Because of this, a standard pulldown experiment for confirmation is not very different from the TAP-tag method and an independent confirmation method is needed. In our opinion, proving an *in vivo* interaction using FRET or split-GFP methods is the best method for confirming this interaction. However, pulldowns using independent reagents (antibodies and protein A or G) might be a satisfactory confirmation.

Trouble Shootings

IgG binding Is Not Efficient as Measured by Western Blot Analysis of Supernatant and Bound Beads. Use 200 μ L of IgG beads per 50 mL protein extract and use

2–4 mL of extraction buffer per gram of leaf tissue. If protein extract is not clear enough, then binding sites are blocked with the nonspecific contaminants; thus try to clear out the extract by either longer period of centrifuge or passing through a non-protein-binding filter. However, it is possible that by lowering the salt concentration in protein extraction buffer, proportion of weakly interacting proteins can be increased, but it will also increase the level of nonspecific binding of proteins to the beads.

Low Recovery after TEV Digestion Using an Antibody to the CBP Domain or Tagged Protein. IgG beads are sticky for many proteins, so elution of proteins even after efficient cleavage of the TAP-tag can be difficult. We have not found an efficient nondenaturing method for eluting these stuck proteins. One possibility that we have not explored is reversing the ordering of the binding steps and eluting the TEV-cleaved proteins from the IgG beads with low pH buffers as is done for protein A:IgG complexes. The protein complex can dissociate at this stage because it will be precipitated and run on an SDS-PAGE gel next. Washing the beads with an additional 200–500 μL of TEV cleavage buffer could recover some eluted protein that is in the dead volume of the beads.

CaM Binding Is Not Efficient. During CaM binding step titration of EDTA coming from the TEV cleavage buffer in IgG eluate is crucial. EDTA is Ca^{2+} chelater and can reduce the CaM-CBP binding efficiency. For this purpose add 3 μL CaCl_2 1 M per mL of IgG eluate.

CaM Elution Is Not Efficient. However, this step is very easy and straightforward and rarely had problem. But if one has a serious problem(s), he/she can think of boiling the bound-CaM beads directly into 2X SDS-loading buffer for 5 min to release the bound protein complex. But remember, boiling the beads is not a specific elution and may give you some noise in the protein interaction data.

36.4 CONCLUSIONS

Protein interaction technologies are essential for studying protein functions in the post-genomics era. TAP-tagging is only a partial solution in plants due to the 11% success rate of obtaining good (stoichiometric) protein complexes. Another 50% of the tagged proteins provide substoichiometric interaction data that supports but does not prove the interactions. This is due to the very low recovery of the putative interacting proteins that are recovered at levels similar to those of possible contaminants. These putative interactors will require additional research to verify their interaction. When the reproducibility of the TAP was tested, it was found to be 70% when using the same TAP-tagged protein complex. The reproducibility is highly influenced by the amounts of the protein complex recovered, with the more abundant proteins being highly reproducible and the lower-abundance proteins being less reproducible. Given the convenient affinity tags and plant transformation systems available, it is worth

trying the method. If the transgenic protein is abundant in the plant, it is worth proceeding with the purification. However, if the expression levels are very low or the tagged protein is recovered without any interacting proteins, it is recommended to not spend large amounts of time and effort to change this outcome. High-affinity stable protein complexes have a fairly high efficiency of recovery while protein complexes with weak interactions are poorly recovered by this method.

36.5 FIVE-YEAR VIEWPOINT

The knowledge of the PPIs provides important information on biological functions of a gene and biochemical mechanisms of the encoded protein. Various protein interaction methods outlined in this book and elsewhere are likely to emerge as important tools to map PPIs. The TAP methodology is appealing because it is a generic purification method for determining PPIs *in vivo*. At the moment, it is the most proven method and offers an important tool for such studies in plants, but methods with higher success rates of purifying intact protein complexes are needed for more efficient progress in this important area.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yinong Yang for his critical comments in the preparation of this chapter. This work was supported by funding from the Nebraska Research Initiative and the NSF (DBI-0217312) to MEF.

REFERENCES

1. IRGSP. 2005. *Nature* **436**:793–800.
2. Tyagi AK and Mohanty A. 2000. *Plant Sci* **158**:1–18.
3. Sasaki T. 2005. The complete rice genome sequence and its application to breeding and genetics, In: *Rice is Life: Scientific Perspectives for the 21st Century*, International Rice Research Institute and Tsukuba (Japan), Los Banos, Philippines, pp. 66–69.
4. Agrawal GK, et al. 2006. *Proteomics* **6**:5549–5576.
5. Junttila MR, et al. 2005. *Proteomics* **5**:1199–1203.
6. Gavin AC, et al. 2002. *Nature* **415**:141–147.
7. Ho Y, et al. 2002. *Nature* **415**:180–183.
8. Forler D, et al. 2003. *Nat Biotechnol* **21**:89–92.
9. Knuesel M, et al. 2003. *Mol Cell Proteomics* **2**:1225–1233.
10. Gully D, et al. 2003. *FEBS Lett* **548**:90–96.
11. van Dijk K, et al. 2005. *Plant Cell* **17**:2439–2453.
12. Rohila JS, et al. 2004. *Plant J* **38**:172–181.
13. Rohila JS, et al. 2006. *Plant J* **46**:1–13.
14. Rubio V, et al. 2005. *Plant J* **41**:767–778.

15. Rohila JS, et al. unpublished results.
16. Rigaut G, et al. 1999. *Nat Biotechnol* **17**:1030–1032.
17. Drakas R, Prisco M, and Baserga R. 2005. *Proteomics* **5**:132–137.
18. Honey S, et al. 2001. *Nucleic Acids Res* **29**:E24.
19. Wydro M, Kozubek E, and Lehmann P. 2006. *Acta Biochim Pol* **53**:289–298.
20. Tanaka A, et al. 1990. *Nucleic Acids Res* **18**:6767–6770.
21. Brockmann B, et al. 2001. *Plant Cell Physiol* **42**:942–951.
22. Witte CP, et al. 2004. *Plant Mol Biol* **55**:135–147.
23. Rivas S, Romeis T, and Jones JD. 2002. *Plant Cell* **14**:689–702.
24. Safadi F, Reddy VS, and Reddy AS. 2000. *J Biol Chem* **275**:35457–35470.
25. Berggren K, et al. 2000. *Electrophoresis* **21**:2509–2521.
26. Gould KL, et al. 2004. *Methods* **33**:239–244.
27. Barrett AJ, Rawlings ND, and Woessner JF. 2003. *Handbook of Proteolytic Enzymes*, Elsevier Academic Press, San Diego, CA.
28. Kayser JP, Vallet JL, and Cerny RL. 2004. *J Biomol Tech* **15**:285–295.
29. Ding X, et al. 2007. A rice kinase–protein interaction Map. Submitted to *Plant Cell*.
30. Ranish JA, et al. 2003. *Nat Genet* **33**:349–355.

TAP STRATEGY IN *ARABIDOPSIS* PROTEIN COMPLEX ISOLATION

Vicente Rubio and Xing Wang Deng

37.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Recent application of TAP-based strategies to protein complex isolation from plants may have provoked a great interest in those among us dealing with the characterization of a particular plant protein. This is because, contrary to tedious classical biochemical purification procedures, the simplicity, high yield, speed, and reliability of the TAP strategy, combined with recent improvements in MS techniques, make it an ideal strategy for protein complex component identification [1]. Moreover, use of the TAP system on a high-throughput scale is already a reality, as shown by proteomic studies using budding yeast [2]. In plants, although at a lower scale, processing of a large number of samples has also been accomplished, as shown by a recent study describing purification of 23 kinase-containing complexes from rice [3] (see also Chapter 36). In addition to *in vivo* interaction studies and protein complex composition analysis, the TAP system permits protein isolation under native conditions, therefore allowing functional studies in which the activity or PTMs on one protein or protein complex can be examined [4].

For TAP isolation of a particular protein, and potentially of its interacting proteins, we first need to fuse in-frame our protein of interest to a TAP tag. The latter

comprises two affinity components surrounding a protease cleavage site. The newly tagged protein can be thus expressed in a particular system or organism and purified using the TAP procedure. This involves basically two different affinity purification stages separated by a specific proteolytic cleavage step. In the original version, used by Rigaut et al. [1] for protein complex isolation from budding yeast, the TAP tag contained two repeats of the IgG-binding motif of *S. aureus* protein A (2xIgG-BD) followed by the cleavage site of the TEV protease, and a CBP that was placed close to the fusion, or proximal region, of the TAP tag (Figure 37.1A). According to this tag composition, total protein extracts containing the corresponding TAP fusion are first incubated with IgG-coated beads (Figure 37.1B). Protein elution from the beads is then performed by cleavage using TEV protease. The second affinity purification step consists in the incubation of the eluted proteins with CaM-coated beads in the presence of calcium. After washing, purified proteins can be recovered by incubating the CaM matrix with EGTA, which in turn impairs CBP binding to CaM by sequestering calcium from the media [1, 4]. The original TAP tag has been successfully used for individual protein isolation from plants, although purification of protein complexes has not yet been accomplished [5].

Starting from this original version of the TAP system, two different alternative tags have recently been generated and applied to plant protein complex isolation. In general, all modifications in these modified versions aim to increase the expression of the TAP-tagged proteins in plants, to improve the yield of purified proteins, to stabilize potential TAP-tagged protein complexes, or to favor proper subcellular localization and functionality of the fusion proteins. Modifications comprise affinity component substitutions and cleavage site replacement. In addition, other changes have been made in the TAP tag-containing vectors that allow easy cloning of target genes into the TAP containing vectors or an increase in TAP fusion expression in plants. Information about all versions of TAP used for plant protein purification is resumed in Table 37.1.

A first modified version of the TAP tag, used for plant protein purification, has the same tag composition as the original one [6]. However, several modifications have been made in the TAP tag sequence to reduce gene fusion silencing or recombination in plants and bacteria, respectively. In addition, a plant nuclear localization signal was found and mutated to avoid fusion proteins to be misplaced within the plant cellular compartments. The authors termed this version as TAPi and have used it to purify fusion proteins that were both transiently and stably expressed in tobacco leaves and rice plants, respectively [3, 6]. Another version, termed TAPa, has been also designed to adapt and improve the TAP procedure for plant protein purification [7]. In the TAPa method, as in the previous one, the first affinity step involves IgG-binding (Figure 37.1B). However, for the second affinity step, the original CaM-binding domain has been replaced with a tag containing a 6 His repeat (6xHis) and 9 repeats of the myc epitope (9xmyc), so that metal affinity and/or anti-myc immunoaffinity chromatography may be performed (Figure 37.1A). Between the two affinity tags, there is a cleavage site for lower temperature active 3 C protease from human rhinovirus, which allows the purification procedure to be performed at 4°C throughout, thus helping to maintain protein stability. In case there is no specific antibody available against our protein of interest, commercially available anti-myc antibodies can be used to detect

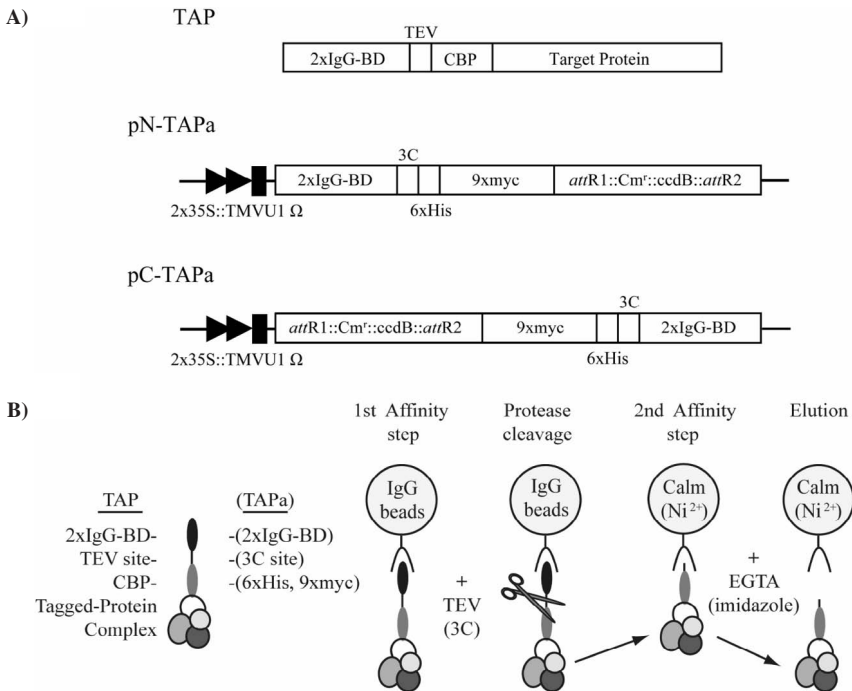


FIGURE 37.1. Comparison between TAP and TAPa tag composition and purification procedures. **(A)** Diagrams representing the expression cassette in TAP and TAPa vectors. Plant-adapted TAP cassette consists of two copies of the protein A IgG-binding domain (2xIgG-BD), a cleavage site for the tobacco etch virus protease (TEV), and a CaM-binding peptide (CBP). The TAPa tag consists of 2xIgG-BD, a 3 C protease cleavage site (3 C), a 6-histidine stretch (6xHis), and 9 repeats of the myc epitope (9xmyc). The pN-TAPa and pC-TAPa vectors allow translational fusion—N-terminal and C-terminal, respectively—of proteins of interest to the TAPa tag. The expression is driven by two copies of the cauliflower mosaic virus 35 S promoter (2x35 S) and a TMV U1Ω translational enhancer. A GATEWAY cloning site (*attR1::Cm^r::ccdB::attR2*) is present in both pN-TAPa and pC-TAPa vectors. **(B)** Schematic representation of the TAP and TAPa procedures. In both cases, purification procedure consists of four different stages: (1) Protein extracts are incubated with IgG beads in both TAP and TAPa systems. (2) Elution of tagged proteins involves specific cleavage on the tag by the TEV protease (TEV; TAP) or the low temperature active rhinovirus 3 C protease (3 C; TAPa). (3) IgG bead eluates are incubated with CaM beads (Calm; TAP) or Ni beads (Ni²⁺; TAPa). (4) Elution of proteins from beads using EGTA (TAP) or imidazole (TAPa). Terms within parentheses in the figure refer to TAPa tag composition and alternative reagents used during TAPa procedure.

our protein fusion at all purification stages. The TAPa tag has been used to generate *Arabidopsis* transgenic lines harboring both N-terminal and C-terminal TAPa fusions of many different light signaling pathway regulators. Molecular characterization of these transgenic lines showed fusion expression in 88% of the genes analyzed and that this expression is largely independent of the fusion orientation. In addition, mutant

TABLE 37.1. TAP Systems Available for Plant Protein Complex Isolation

System ^a [Ref.]	Vector's Name	Tag Composition ^b	Availability ^c	Comments
TAPi [3, 6]	CTAPi NTAPi	2xIgG-BD, TEV, mCBP	Upon request from the authors	Plant adapted
TAPa [7, 8]	pC-TAPa pN-TAPa	2xIgG-BD, 3 C, 6xHis, 9xmyc	ABRC; CD3-679 ABRC; CD3-696	Plant adapted
sTAPa [7]	pVR-TAP Ct pVR-TAP Nt	2xIgG-BD, 3 C, 6xHis	Upon request from the authors	Plant adapted
pEarlyGate [20]	pEarlyGate-205	2xIgG-BD, TEV, CBP	Upon request from the authors	Fusion may localize mainly in the nucleus

^a All systems contain Gateway-based cloning sites. sTAPa: smaller variant of TAPa.
^b IgG-BD, IgG-binding domain of protein A; TEV, TEV protease cleavage site; mCBP, mutated CBP; 3 C, 3 C protease cleavage site
^c Stocks numbers are provided if available. ABRC, *Arabidopsis* Biological Resource Center at the Ohio State University.

complementation analysis showed that most of TAPa fusions analyzed retained function of the wild-type proteins. The suitability of the TAPa system to allow efficient multiprotein complex isolation from stably transformed *Arabidopsis* was confirmed by purifying all eight components of the COP9 complex (CSN) from plants expressing a C-terminal fusion between the TAPa tag and one of CSN's subunits, CSN3 [7, 8]. A smaller variant of the TAPa tag (sTAPa) has been obtained by removing the nine repeats of the myc epitope. The resulting tag (18 kDa) may overcome potential interference with complex assembly or protein function due to the relatively large size (33 kDa) of the TAPa tag [7].

The major goal of this chapter is to help neophytes to design and adapt the TAP strategy towards purification of their protein(s) of interest. Although this chapter will mainly be focused in the TAPa method applied to protein isolation from *Arabidopsis*, many of the comments and ideas herein can be extended to all TAP versions used for protein purification from any plant species. As well, we provide some tips and advice to help TAP users to improve fusion protein yield and quality and to overcome problems while expressing or TAP-purifying proteins from plants.

37.2 SPECIFIC METHODOLOGY AND STRATEGIES

Making Fusions

Generation of TAP fusions is highly conditioned to the preliminary information available concerning our protein(s) of interest. In this regard, any reported purification attempt or preceding characterization of our target protein(s) or related proteins can shed light on specific molecular features that should be taken into account prior to generation of the corresponding TAP fusions. Thus, it has been reported that maintenance

of a particular fusion orientation, expression levels, or use of a specific tag is very often essential for proper protein complex assembly, localization and function. This is the case of *Arabidopsis* E3 Ub-protein ligase COP1, for which addition of a peptide at its C-terminus causes loss of function by interfering with a WD repeat-containing domain used for substrate interaction [9]. Use of truncated versions has also been revealed as a prerequisite to successfully express and purify certain proteins. We recommend that, in case no previous data concerning our protein(s) of interest is available, different alternatives should be tested, such as generation of both C-terminal and N-terminal TAP fusions, constitutive overexpression and endogenous promoter-driven expression of fusions, and TAP-tagging of both full-length and truncated versions of target protein(s). Most likely, all these alternatives cannot be taken into consideration when we are addressing purification of a large number of proteins but when tagging a limited number of them. Nevertheless, generation of multiple TAP constructs for purification of a large set of proteins or numerous variants of the same one can be facilitated by cloning procedures based on site-specific recombination, such as Gateway (Invitrogen) and Cre-lox technologies [10]. Vectors for expression of TAP-tagged proteins in plants include Gateway based-recombination sites in their N-terminal and C-terminal expression cassettes (Table 37.1). Although the conventional Gateway reading frame is commonly used, one should be aware that in the TAPa (and sTAPa) system the reading frame has been modified [3, 7]. This different reading frame is only relevant when generating N-terminal fusions (tag-target gene), but not when preparing C-terminal fusions (target gene-tag) for which the conventional reading frame is still used.

The conventional Gateway reading frame is described in *Section 2.3.2 “Reading Frame”* for Gateway cloning technology: “For fusion proteins, it is essential to establish the correct reading frame. For N-terminal fusions, construct the Entry Clone so that the DNA sequence is in frame with the two Lys codons (AAA–AAA) found in *attL1* (or *attB1* for PCR primer design). For C-terminal fusions, construct Entry Clones so that the DNA sequence is in frame with the Tyr-Lys (TAC–AAA) in *attL2*”. For cloning in the N-terminal TAPa and sTAPa vectors (Table 37.1), instead of the two Lys codons, the DNA sequences in the Entry Clones must be in frame with the AAA-AGC sequence present in *attL1*. Below is the DNA sequence for the *attB*-PCR primers used to amplify and clone genes of interest into the Entry vector by the BP reaction (see Gateway manual) so that they are suitable for both TAPa (and sTAPa) N- and C-terminal fusion:

***attB*-PCR primer. Forward:**

5'-GGGGACAAGTTTGTACAA AAAAGC AGG CTA TAC AAA **atg** (1st Met)
+20 to 27 nt from the gene of interest –3'

***attB*-PCR primer. Reverse:**

5'-GGGGACCAC TTTGTA CAA GAA AGC TGG GTA **tta** (optional STOP
codon) +20 to 27 nt from the gene of interest –3'

Please note that when making N-terminal TAPa fusions, if the Entry clone does not contain a STOP codon, 14 extra amino acid residues will be added to the final

fusion protein at its C-terminus. Since the empty TAPa (and STAPa) vectors, as all Gateway technology-containing plasmids, contain the *ccdB* gene, they must be transformed and propagated into a bacterial strain that is resistant to the lethal effects of CcdB protein, such as the DB3.1 competent cells (Invitrogen). For propagation of both empty and fusion containing plasmids, the bacterial resistance gene is for spectinomycin (50 mg/L). Notice that the chloramphenicol resistance gene (25 mg/L) located between the *attR1* and *attR2* regions of the Gateway cloning cassette can be used only during propagation of the empty plasmids. If TAPa transgenes are stably transformed into *Arabidopsis* plants, transgenic lines selection should be performed in gentamicin-containing medium (gentamicin sulphate, 200 mg/L). In some genetic backgrounds, it might be necessary to reduce the antibiotic concentration (150 mg/L). Although transient expression of TAPa fusions has been successfully used to purify protein complexes from tobacco leaves, such an approach does not allow molecular-functional characterization of the TAP fusion in a whole plant context. To avoid this limitation, we suggest use of stably transformed transgenic plants where, in addition, the effects on the protein complex composition, modifications, and activity due to a wider range of growth conditions, stimuli, and plant material used during purification could be assessed [7].

Upon plant transformation, two copies of the constitutive 35 S promoter of the CaMV drive expression of TAPa transgenes, leading to overexpression of the corresponding recombinant proteins. Overexpression of the target protein has been detracted since it might lead to the formation of non-natural or nonspecific interactions, thus compromising identification of the structure, composition, and/or activity of targeted protein complexes [4]. Nevertheless, overexpression of protein fusions should also increase tagged protein incorporation into its corresponding protein complex in those cases in which our protein of interest is the limiting factor for complex formation. Additionally, in case a mutant or silencing line for a specific gene is not available, overexpression of the tagged protein should help tagged protein incorporation into the protein complex in preference to the endogenous protein [7].

Finally, studies using both the TAPa and the TAPi system have shown that numerous nonspecific proteins appear in purified samples [3, 6–8]. Thus, we recommend use of transgenic lines containing TAP fusions to unrelated proteins as negative controls. Plants expressing TAP fusions to exogenous proteins, such as GUS or GFP, can help to discern specific interacting proteins in the purified samples. In this regard, a TAPa-GFP containing line is available upon request to the authors of this chapter.

Functional Characterization of Transgenic Lines

Prior to purification, a preliminary functional and molecular characterization of the TAPa-tagged proteins should be performed. For this, independent *Arabidopsis* transgenic lines for each TAPa fusion should be obtained and fusion protein levels for each independent line should be compared using immunoblots against the myc epitope. Fusion expression levels may vary greatly among independent transgenic lines for the same TAPa fusion possibly reflecting positional effects at transcriptional level due to the transgene location [7]. Although one major band corresponding to the

TAPa-tagged protein should be expected, previous results have shown that other specific minor bands might appear, probably corresponding to truncated versions of the corresponding TAPa fusion. Indeed, additional bands used to correlate with a high level of transgene expression, suggesting that the higher the expression is, the more likely it is that the TAPa fusions are processed and truncated by nonspecific proteases. In addition, for some specific tagged proteins, fusion size might look much bigger than expected, according to its electrophoretic mobility. This anomalous electrophoretic behavior might be due to a wide variety of PTMs, such as glycosylation, ubiquitination, and sumoylation, among others [11]. On the other hand, slower mobility in denaturing gels has been shown to be inherent to specific target proteins [12, 13]. Isolation and determination of each band composition, upon TAPa procedure, should help to discriminate between these two possibilities.

Functional characterization of transgenic lines can be accomplished by performing a mutant phenotype complementation analysis. In the case where the TAPa transgenes were not transformed into the corresponding mutant background, study of potential gain-of-function phenotypes due to overexpression of the targeted proteins can provide insights into their functionality. In both cases, transgenic plants should be grown in the specific conditions reported for each genetic background and different parameters should be analyzed and compared to those in mutant and/or wild-type plants grown in the same conditions. The degree of complementation may vary, depending on the growth conditions, as shown during complementation analysis of the *hy5* mutation using N-terminal *TAPa-HY5* transgenic lines [7]. Thus, plants expressing TAPa-HY5 fully complemented the mutant background when grown under continuous red or blue light. However, the same transgenic lines grown under continuous far red light showed a partial complementation. For fusion protein purification purposes, proteins should be preferentially extracted from transgenic plants grown in the specific conditions where full complementation or clear gain-of-function phenotypes are observed. Additionally, fusions function can be assessed analyzing any biochemical or biophysical activity inherent to the target protein. As an example, functionality of CSN3 has been analyzed by testing the capacity of the CSN3-TAPa containing-CSN to derubylate cullin 1 *in vitro* and *in vivo* [7, 8]. Finally, if possible, gel filtration assays should be performed to determine whether TAPa fusions are correctly incorporated into endogenous protein complexes. In this regard, the gel filtration profile of the TAP fusion should be similar to that of the endogenous target protein, although a size difference between their peak fractions may occur, possibly due to the presence of the TAPa tag.

Preparing the Plant Material

It has been previously mentioned that fusion proteins should be isolated preferentially from transgenic plants cultivated under specific growth conditions (temperature, luminosity, humidity, media composition, etc.) in which full functionality of tagged proteins has been confirmed. In this scenario, we will be able to assume that regulation of the biochemical activity, map of protein interactions, and subcellular localization of a particular fusion resembles that of the endogenous protein. Once growth conditions for a specific transgenic line have been chosen, accumulation of the recombinant protein should be assessed in a time course experiment. A compromise ought to

be reached between peak accumulation of tagged proteins and amount of fresh tissue. Thus, it is possible that a particular tagged protein accumulates at its highest in 5-day-old seedlings and only a half of these levels remain after two weeks. However, 2-week-old seedlings may provide about 10 times more fresh material than 5-day-old plants, thus containing a higher amount of total tagged protein.

Regarding the amount of fresh plant material necessary for successful TAPa purification, it generally depends on the level of expression of each tagged protein. Our experience has shown us that purification of silver staining-detectable levels of a multimeric protein complex can be accomplished by using just 1.5 g of *Arabidopsis* fresh tissue [7, 8]. However, this is an optimal situation and, in general, more than 10 g plant material is necessary to obtain MS detectable levels of tagged protein.

TAP Protocol for *Arabidopsis* Protein Complex Purification

The protocol described below has been optimized for purification of tagged-proteins from *Arabidopsis*, and it is fully useful only when using the TAPa system. Note that the second affinity purification stage in the TAPa protocol herein involves incubation with Ni-coated beads, and there is not any reference to a potential immunoaffinity purification step using anti-myc beads. In our experience the second possibility yields less purified fusion proteins and requires expensive or difficult-to-prepare reagents. Thus, use of anti-myc beads is not recommended unless metal affinity purification is unsuitable for our purposes. In this case, commercially available anti-myc-coated beads (Sigma, Covance) might represent a suitable alternative.

1. Harvest approximately 15 g plant tissue for both sample and negative control. The amount may vary, depending of the level of expression of the corresponding TAPa fusion. Grind the plant material in liquid nitrogen using a mortar and a pestle. Thaw the homogenate in 2 volumes of extraction buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% NP-40, 1 mM PMSF, and 1X protease inhibitor cocktail [Roche]), filter it through four layers of cheese-cloth, and centrifuge at 12,000g for 15 min at 4°C. The supernatant is the total extract.
2. Incubate the total protein extract for 2 h at 4°C with 500 µL IgG beads (Amersham Biosciences) that has been pre-washed 3 times with 10 mL extraction buffer. Wash the beads 3 times with 10 mL of washing buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% NP-40) and once with 10 mL of cleavage buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% NP-40, 1 mM DTT).
3. Elute proteins from the IgG beads by incubation with 50 µL (100 units) of 3 C protease (Prescission protease, Amersham Biosciences) for 2 h in 5 mL of cleavage buffer at 4°C with gentle rotation. Perform another elution and pool with the first eluate.
4. Load the entire 10-mL eluate onto a 1-mL Ni-NTA resin column (Qiagen). Reload the flow-through onto the column once. Wash with 30 mL of washing buffer.

5. Elute proteins from the Ni-NTA resin using 1 mL of imidazole elution buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% NP-40, 0.05 M imidazole).
6. Load 50- μ L eluted proteins onto 4–20% gradient denaturing gels and separate by SDS-PAGE. Determine purity and yield of isolated proteins by silver staining.
7. For protein identity determination, stained protein bands can be excised and subjected to MS analysis. Note that MS techniques may require compatible protein staining procedures (among others, silver staining protocols where glutaraldehyde is omitted, CBB stain, and SRPGS [Invitrogen]). Alternatively, eluted proteins can be analyzed in bulk—for example, by MudPIT [14]. Bioinformatic comparison of the resulting peptides in each sample to those predicted for *Arabidopsis* proteins will allow protein identification.
8. For biochemical activity assays, eluted samples can be dialyzed when imidazole needs to be removed. Alternatively, imidazole-containing buffer can be changed into buffer of choice using desalting spin columns (Pierce). Samples can be concentrated using microcon YM-3 filter devices (Millipore).

Improving Protein Complex Recovery and Purity

Upon TAP purification of target protein complexes, high levels of contaminant proteins and/or low protein recovery may occur. To cope with these problems, several corrective actions can be taken. Regarding protein contaminants removal, these generally correspond to proteins from the reagents employed during TAPa purification (i.e., 3 C protease and IgG) and to nonspecific interactors that interact with the TAPa tag or with the beads used. In the first case, incubation of beads with acid solution can help to decrease the amount of free IgG in the IgG-coated resin. Reduction of the 3 C amount can be carried out using glutathione sepharose since the 3 C protease is commercially available as a fusion fused to the GST tag. In the case of nonspecific interacting proteins, increasing the stringency of the procedure, by using higher concentrations of salt and detergent in the buffer solutions, should help to reduce their level [7]. If protein contamination persists, additional purification steps can help such gel filtration procedures. Moreover, the presence of three different tags (2xIgG-BD, 6xHis, and 9xmyc) in the TAPa system could allow a triple affinity purification procedure, although this has yet to be demonstrated. Keratins frequently represent a common source of sample contamination. Avoid keratins contamination by using gloves throughout TAPa procedure but especially when excising protein bands (at this point use of a mask is highly recommendable). Also, the wearing of woolen clothes during samples manipulation should be avoided.

If greater amounts of tagged protein or protein complex are required, scaling up should be performed by processing of 6–10 identical samples in parallel, instead of a single large TAPa purification. This recommendation is in agreement with previous results obtained in TAP purification assays applied to yeast and plant protein complex isolation [4, 7]. At the last step of the TAPa procedure, Ni beads eluates can be pooled, and proteins therein can be concentrated and separated by SDS-PAGE. Additionally,

cross-linking may help to stabilize weak interactions within protein complexes, leading to an increase in the purification yield and purified proteins number as it has been previously reported [3]. However, one should be aware that it may also potentially difficult TAPa fusion binding to the IgG-coated beads or cleavage by the specific protease by hiding the corresponding TAPa tag domains. The suitability of cross-linking for TAPa-tagged protein complex isolation will depend thus on the specific molecular features of each independent TAPa fusion. A protocol for protein crosslinking using *Arabidopsis* fresh tissue is described below (modified from Rohila et al. [3]). All steps should be carried out in a cold room:

1. Vacuum-infiltrate fresh plant material with ice-cold 1X PBS and 1% formaldehyde (15 g fresh plant material/100 mL solution) during 30 min.
2. Rinse plants with 100 mL ice-cold 300 mM glycine (aqueous solution).
3. Vacuum-infiltrate the plant material with 100 mL ice-cold 300 mM glycine (aqueous solution) during 30 min.
4. Rinse plants twice with 100 mL ice-cold 1X PBS.
5. Carefully dry the plant material in paper napkins, thereby avoiding plant tissue breakage.
6. Freeze in liquid nitrogen. Continue with the TAPa procedure.

37.3 EXPERIMENTAL RESULTS AND APPLICATIONS

As part of our long-time running interest in the regulation of plant development by light, we have studied the mechanisms controlling HY5 function. The latter is required for activating light-regulated gene expression and is also a target of the E3 Ub-protein ligase activity of COP1, a known repressor of photomorphogenesis under darkness [13]. Previous results show that HY5 functions exclusively as a protein complex *in vivo* [15]. To proceed with purification of the HY5 complex, we took advantage of the TAPa system. For this, *Arabidopsis hy5* transgenic lines overexpressing TAPa fusions to the HY5 N terminus (TAPa-HY5) and C terminus (HY5-TAPa) were obtained. Complementation analysis of the *hy5* phenotype in the resulting transgenic lines showed a partially complemented phenotype and an almost completely complemented phenotype for the C-terminal and N-terminal lines, respectively. Preliminary gel filtration data revealed the presence of a putative HY5 fusion-containing complex in protein extracts of these transgenic lines (Figure 37.2A). These assays also showed that the size of this complex was bigger (~400 and ~280 kDa for the C- and N-terminal fusion, respectively) than the expected for HY5 fused to TAPa (~230 kDa). The difference in size between both C- and N-terminal fusions could be due to an interaction of the TAP tag with nonspecific plant protein(s) as previously reported [2]. Furthermore, nonspecific interacting protein(s) could be impairing proper tagged HY5 complex functionality, mostly observed in transgenic lines harboring the C-terminal fusion. According to gel filtration assays and complementation of *hy5* phenotype results, the TAPa-HY5 expressing lines were chosen for further analysis. At the time these experiments were

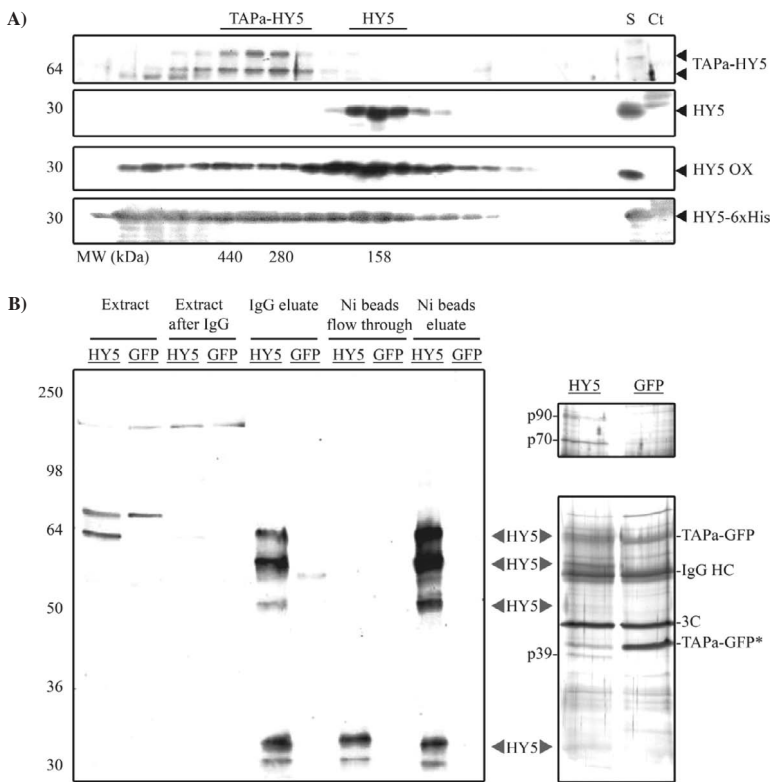


FIGURE 37.2. Molecular characterization and purification of TAPa-HY5. **(A)** Immunoblot analysis of gel filtration protein fractions from wild-type (HY5), TAPa-HY5-expressing (TAPa-HY5), and HY5-overexpressing (HY5 OX) lines. Each line protein extract was used as a positive control (S). Negative controls (Ct) correspond to *hy5* protein extracts. The profile for a bacterial expressed HY5 fusion (HY5-6xHis) is also shown. The α -HY5 Ab was used in all cases. Black arrowheads point to the corresponding HY5 version. For TAPa-HY5, two bands are observed, as previously reported [7]. The positions of the MW standards are labeled at the bottom. The peak fractions containing the endogenous HY5 and TAPa-HY5 complexes are indicated. **(B) Left panel:** Immunoblot of different fractions obtained during the TAPa-HY5 purification. Protein extracts from *TAPa-HY5* (HY5) and *TAPa-GFP* (GFP, negative control) plants were subjected to TAPa purification. Proteins corresponding to each fraction obtained were separated on a 10% SDS-PAGE. The α -HY5 antibody was used for immunoblotting. Gray arrowheads point to full-length and truncated versions of TAPa-HY5 fusion. The following fractions are shown; Protein extract (Extract, 0.1%), unbound protein after IgG beads incubation (Extract after IgG, 0.1%), eluate from IgG beads using 3 C protease (IgG eluate, 1%), discarded Ni-NTA column flow through fraction (Ni beads flow through, 1%), and final purified protein (Ni beads eluate, 5%). Percentages indicate the amount of protein loaded relative to the total amount of protein in each sample. **Right panels:** Visualization of purified TAPa-HY5 and TAPa-GFP protein bands. 95% of the Ni beads eluates were separated on a 10% SDS-PAGE and used for silver staining. The position of three TAPa-HY5 interacting proteins (*upper panel*: p90, p70; *lower panel*: p39), the IgG heavy chain (IgG HC), 3 C protease (3 C), and both full-length (TAPa-GFP) and truncated (TAPa-GFP*) is indicated.

carried out, it was reported that, in some cases, the large apparent size observed for a putative protein complex by gel filtration assay depends on intrinsic properties of the protein sample, rather than being caused by association with HMW proteins [16]. We checked if this was the case for the HY5 complex. In this context, we performed gel filtration assays by using protein extracts from a recombinant HY5-expressing bacterial strain (Figure 37.2A). The resulting gel filtration profile was very similar to that obtained when using plant protein extracts, suggesting that HY5 itself causes its large apparent size by gel filtration. This result does not discount the presence of small proteins other than HY5 in the HY5 complex, but it lowers the likelihood that large proteins are HY5 complex constituents.

To identify proteins that are either stable constituents of the HY5 complex or transient HY5 complex interactors, a low-stringency TAPa procedure was applied to protein extracts from both TAPa-HY5 and TAPa-GFP (negative control) expressing lines (Figure 37.2B). As a result, we could observe by silver staining three differential protein bands (referred to by their MW as p90, p70, and p39) that co-purified with TAPa-HY5, but were not present in the negative control (Figure 37.2B). Analysis of the sliced gel bands by MS revealed the identity of several candidate proteins as putative HY5 interactors. Among them, SPA1-like protein 3 (SPA3) and a light-repressible receptor protein kinase, corresponding to the p90 and p70 proteins, respectively, have been already related to the light-signaling pathway, but their roles have not been completely elucidated [17, 18]. It has been suggested that the SPA proteins are core constituents of a COP1-containing complex and regulate COP1's E3 Ub-protein ligase activity toward positive regulators of photomorphogenesis, such as HY5 [9]. For p39, there is no homology to any other proteins sequence. The study of this novel protein, in addition to the further characterization of the reported ones, represents a major goal in order to gain new insights in the HY5-mediated light signal transduction.

Our current model depicts a functionally active HY5 complex containing fewer proteins than suggested previously (possibly, it might be formed just by a HY5 homodimer), with which some other proteins, such as SPA3, interact transiently in order to finely regulate its activity.

37.4 CONCLUSIONS

The number of studies that have been published in the plant proteomics area has increased “notoriously” during the last three years. However, this trend may represent more a promise than a reality because many of these studies are reviews and opinion articles about this area, rather than original articles [19]. The need of adaptation of the proteomics tools, widely used in other model organisms, to the particularities of plant systems seems to greatly substantiate this fact. The application of TAP-based strategies for plant protein complex isolation may help to overcome this limitation. Several plant-adapted TAP procedures are now available to the plant science community [3, 5, 7, 20]. Among them, the TAPa system has been proven an efficient alternative to other affinity purification methods and classical biochemical protein purification procedures. TAPa allows multiprotein complex isolation from both stably and transiently

transformed-plant material. Indeed, it has been used to purify protein from different plant species, such as *Arabidopsis* and tobacco. Previous studies have already shown that protein complex samples obtained by using the TAPa strategy are suitable for MS analysis purposes and biochemical activity assays [7, 8].

37.5 FIVE-YEAR VIEWPOINT

Several improvements in TAP methodology should be expected in the near future, helping TAP-based technologies to become widely used within the plant research community. Although purification of a large number of plant protein complexes using TAP has been already carried out, in most cases, sample preparation still needs to be optimized for each sample type. Thus, standardization of protocols, allowing reduction of the amount of starting material needed and high-throughput approaches in plants, will truly enable use of TAP as a common tool in plant biology laboratories. At the technical level, improvements will include use of smaller tags composed by shorter-affinity domains, such as Flag, HA and streptavidin motifs. Also in order to reduce costs and sample contamination due to exogenous proteases, specific cleavage should not be a requirement anymore. Instead, high-affinity competitors for the corresponding TAP components, such as Flag, HA, or avidin peptides, will represent a very convenient alternative. Additionally, TAP procedures should allow protein isolation at reasonable levels of purity using just one affinity step. Thus, minimal loss of samples could be accomplished in those cases in which fusion expression is very low or plant material is limited. In this regard, promising results have already been obtained using an HA-streptavidin II tag [21]. In addition, vectors allowing inducible expression and/or cell-type or tissue-specific expression of TAP transgenes should be available within the next five years. Finally, the range of plant species for which TAP will become an efficient method for protein isolation will increase. However, in most cases this will require that genomic information be first obtained for these species. Obtainment of a standard TAP protocol suitable for any plant species will really boost the application of this technology in plant proteome studies.

REFERENCES

1. Rigaut G, et al. 1999. *Nat Biotechnol* **17**:1030–1032.
2. Gavin AC, et al. 2002. *Nature* **415**:141–147.
3. Rohila JS, et al. 2006. *Plant J* **46**:1–13.
4. Puig O, et al. 2001. *Methods* **24**:218–229.
5. Rivas S, Romeis T and Jones JD. 2002. *Plant Cell* **14**:689–702.
6. Rohila JS, et al. 2004. *Plant J* **38**:172–181.
7. Rubio V, et al. 2005. *Plant J* **41**:767–778.
8. Menon S, et al. 2005. *Methods Enzymol* **398**:468–481.
9. Saijo Y, et al. 2003. *Genes Dev* **17**:2642–2647.
10. Hajdukiewicz PT, Gilbertson L, and Staub JM. 2001. *Plant J* **27**:161–170.

11. Kwon SJ, et al. 2006. *J Exp Bot* **57**:1547–1551.
12. Holm M, et al. 2002. *Genes Dev* **16**:1247–1259.
13. Osterlund MT, et al. 2000. *Nature* **405**:462–466.
14. Chen EI, et al. 2006. *Mol Cell Proteomics* **5**:53–56.
15. Hardtke CS, et al. 2000. *EMBO J* **19**:4997–5006.
16. Van Der Hoorn RA, et al. 2003. *Plant J* **35**:305–315.
17. Laubinger S, Fittinghoff K, and Hoecker U. 2004. *Plant Cell* **16**:2293–2306.
18. Deeken R and Kaldenhoff R. 1997. *Planta* **202**:479–486.
19. Rossignol M, et al. 2006. *Proteomics* **6**:5529–5548.
20. Earley KW, et al. 2006. *Plant J* **45**:616–629.
21. Witte CP, et al. 2004. *Plant Mol Biol* **55**:135–147.

BLUE-NATIVE PAGE IN STUDYING PROTEIN COMPLEXES

Holger Eubel and A. Harvey Millar

38.1 INTRODUCTION

Protein Complexes

The term “proteomics” is an umbrella term under which several different levels of assessment of a given set of proteins are combined. One of these levels is the analysis of the interactions between proteins forming this proteome, also termed “interactomics.” Many types of interactions between proteins exist, ranging from short-lived ones with durations of merely fractions of a second to stable ones whose duration might only be affected by the natural turnover rate of the participating proteins. Because it is not possible to exactly determine where the changeover point occurs from short-lived to reasonably stable interactions, it is important to note that a validation of PPIs is most feasible for the latter type. This type of interaction leads to the formation of what is commonly referred to as a “protein complex.” The presence of many protein complexes in various compartments of plant and non-plant cells is widely accepted today. Protein complexes are known to play important roles in such essential processes as metabolism, gene expression, replication and repair of genetic material, and protein turnover. A wealth of undiscovered protein complexes still awaits us in mitochondria,

plastids, and peroxisomes as well as other compartments of the eukaryotic cell. The advantages of protein complexes over a series of single enzymes are manifold. They include reduced diffusion distances of the substrates within the complex, increased speed of the overall reaction when compared to a series of individual enzymes, and faster processing of (unstable) intermediates, thereby reducing unwanted side reactions and a loss of substrate. Also, the capacity of biological membrane bilayers to take up proteins is increased if these are merged into complexes. Finally, another possible advantage of protein complexes over a cascade of singular enzymes might be found in the establishment of simplified mechanisms to control the metabolite flow through biochemical pathways. Because of the tight association of the proteins within a complex, often a structural interdependence of the proteins can be observed. The modification of a single subunit of the complex might therefore lead to its dissociation or is sufficient to activate or deactivate several other enzymes of this unit. This last point is of special importance in the case of “accessory” subunits in complexes, which are not involved in the primary function of the complex.

In a biological context, study of stable PPIs can generate further knowledge or provide starting points for continuing research employing a diverse range of approaches: (i) If a biochemical pathway consists of a series of complexes in a row, branching of this pathway can only occur between those units. This drives a biochemical pathway to efficiently yield a specific end product, without intermediates being bled off into competing pathways. The association of protein complexes into supercomplexes has been demonstrated impressively by the analysis of the structure of the mitochondrial electron transport chain, facilitated by the use of specific detergents [1]. By analogy to the investigation of protein complexes, the analysis of supercomplexes is likely to have a strong impact on our understanding of the bifurcation and amalgamation of biochemical pathways in the complex metabolism of plants. (ii) An important factor that contributes to this increased complexity of plant metabolism over that of other organisms is the presence of additional subunits in plant protein complexes which are not involved in its primary function. Instead, these subunits introduce side functions into the complexes [2, 3]. The presence of such accessory subunits can indicate an interdependence of the primary function of the complex with another metabolic pathway and the introduction of the additional subunits may have a regulatory role. By ascertaining the identity and function of the accessory subunits, it is possible to determine the nature of this second pathway. To determine whether such subunits actually participate in the formation of a protein complex, the direct isolation of intact complexes is the only feasible way because the pool of potential candidates is very large. (iii) The identification of proteins in the course of several mitochondrial and plastidic proteome projects has led to the discovery of a wealth of proteins with unknown function [4, 5]. Some of these unknown proteins have also been found to be subunits of protein complexes such as the plant respiratory complexes I and II [2, 6]. While the analysis of protein complexes may not necessarily elucidate the function of new proteins, it may be able to put them into a narrow functional context. (iv) In the model plant *Arabidopsis*, paralogous proteins can be found participating in the formation of protein complexes within different organelles. Careful analysis of the protein complex subunits by MS can reveal the distribution of these paralogues within the cell [7].

The finding that a protein found in chloroplasts and mitochondria is assembled into a major protein complex in one of the organelles is a strong indication for a genuine targeting into this particular organelle (see SUBA at www.suba.bcs.uwa.edu.au [8]). (v) The analysis of the subunit composition of a protein complex can also be the starting point for reverse genetic approaches to probe the importance of an accessory protein for the function of the complex as a whole.

A targeted analysis of the promoter elements of the genes coding for subunits of a complex is also possible to understand the mechanisms of co-regulation of the expression of a complex. Again, promoters that drive the expression of accessory plant-specific subunits will be of special interest, because a difference in the expression pattern of those genes may enable us to draw conclusions about additional regulatory mechanisms.

BN-PAGE

In their attempts to study the protein complexes of the respiratory chain of mitochondria, Schägger and von Jagow developed a gel-based system able to separate all of the protein complexes involved in oxidative phosphorylation in their native state [9]. This provided a new method in addition to the commonly employed enzymatic assays, densitometric methods, or chromatographic purification methods. By mixing the sample with CBB before the gel run, they introduced a large number of negative charges into the protein complexes, overshadowing their intrinsic charges. This was the essential step for a complete and sharp separation of the mitochondrial protein complexes. The use of CBB and the native approach were eponymous for this gel system, which consequently has been termed BN-PAGE. Combined with a denaturing SDS-PAGE gel, BN-PAGE forms the first dimension of a 2D system which is able to resolve protein complex subunits according to their affiliation to a certain complex and according to their molecular mass. The typical protein pattern on such a 2D BN/SDS-PAGE gel consists of a sample specific array of vertical rows of proteins, each row representing the subunits of a single type of protein complex (Figure 38.1). Rows on one end of the gel represent protein complexes with an HMW, while those found on the other end belong to complexes with an LMW. Within these rows, proteins with an HMW will be found near the top, while those with an LMW will migrate nearer to the bottom of the gel.

A clear advantage of BN-PAGE compared to other methods lies in its direct analysis of endogenous proteins and its global approach. No genetic manipulation is required. This is not the case for the popular Y2H system, TAP, or the FRET/BRET technology approaches. The Y2H approach, in particular, is known to create a considerable number of FP results, probably due to analysis outside the normal environment of the interactions under investigation. The analysis of each potential PPI using these techniques also requires an independent experiment. BN-PAGE, in contrast, allows the investigation of a multitude of protein interactions by a single experiment and therefore can be considered as a large-scale approach to this question. It also allows an assessment of the MW of the complex and its abundance. Limitations of BN-PAGE compared to the above-mentioned techniques are discussed in Box 38.1.

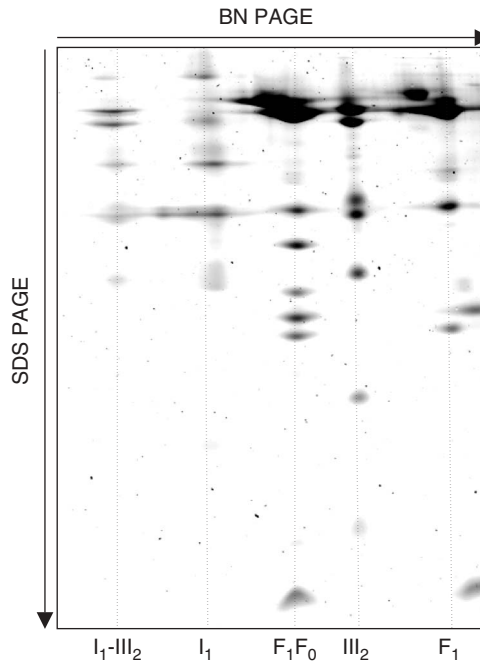


FIGURE 38.1. 2D BN/SDS-PAGE separation of *Arabidopsis* mitochondrial protein complexes. Solubilization of the complexes has been accomplished using digitonin in a ratio of 10 g detergent per g protein prior to BN PAGE. Dotted lines indicate protein complexes resolved into their subunits. Arrows on top and to the left of the gel show the direction of separation in first and second dimension from high to low MW. The identity of protein complexes is given below the gel. **I₁-III₂**: supercomplex consisting of a single copy of NADH dehydrogenase (complex I) and the dimeric version of Cytochrome c reductase (complex III). **I₁**: complex I. **F₁F₀**: mitochondrial ATP synthase (complex V). **III₂**: dimeric complex III. **F₁**: Soluble part of complex V. *Note*: Breakdown of respiratory protein complexes is observed very rarely with the use of digitonin. If it does, it might indicate that the detergent and/or Coomassie concentrations in the sample were too high.

38.2 BRIEF BIBLIOGRAPHIC REVIEW

BN-PAGE was initially developed to study the composition of the respiratory chains of mammals, yeast, and bacteria [9], which remained a focal point until today. It has especially been employed for the analysis of mitochondria-related diseases in humans and only a small number of studies have used BN-PAGE for the analysis of other, nonrespiratory protein complexes. In plants, the first study utilizing BN-PAGE was published in 1994 on the chloroplast cytochrome b(6)f complex [10]. Since then, BN-PAGE in plant proteomics has been dominated by the analysis of electron transfer chain complexes in plastids and mitochondria. From the beginning, *n*-dodecylmaltoside (DDM) and Triton X100 have been the detergents of choice for the solubilization of the complexes. The introduction of digitonin (Box 38.2) for the solubilization of protein

Box 38.1 Limitations of BN-PAGE

While BN-PAGE is a powerful tool to study stable PPIs, there are a range of applications and conditions for which it is not suitable:

1. Temporary interactions between proteins often cannot be analyzed by BN-PAGE. This problem increases as the duration of the interaction decreases.
2. Membrane protein complexes require solubilization prior to BN-PAGE. The success of the gel run depends significantly on the choice of detergent and the establishment of the right mass ratio between detergent and protein. Often, even careful sample preparation cannot avoid dissociation of protein complexes and supercomplexes.
3. Soluble protein complexes (and, to a certain extent also those inserted into membranes) are often destabilized by the use of CBB [30, 47]. If a reduction in the amount of this dye is not improving protein complex stability, CN-PAGE with all of its generic limitations often is the only alternative.
4. In protein complexes containing more than two proteins, BN-PAGE fails to assign which proteins are directly interacting or associated.
5. For samples containing a significantly wider selection of protein complexes of relatively low abundance, subfractionation by other means than molecular size might be necessary. Alternatively, immunoblots have been used successfully to demonstrate the presence of proteins in low-abundance complexes. This, however, is done at the expense of the broadness of the approach.

complexes from the inner mitochondrial membrane of yeast and mammals in 2000 set a new benchmark in the development of this technique [1]. It allowed the discovery of associations of whole respiratory protein complexes into supercomplexes. The first studies on plant mitochondria and chloroplast employing digitonin were conducted in 2004 [11, 12].

38.3 METHODOLOGY AND STRATEGY

BN-PAGE

In many respects, the practicalities of BN-PAGE are similar to a standard SDS-PAGE experiment. However, there are some features that clearly distinguish these two systems from each other. The main feature is the total absence of SDS because of its high solubilization and denaturing power. In normal SDS-PAGE, SDS is involved in the solubilization and denaturation of proteins and in the attachment of negative charges to them. Because a denaturation of the sample is not desired in BN-PAGE, solubilization/denaturation and the addition of charges are accomplished by two different reagents: a mild, non-ionic detergent and the dye CBB.

Box 38.2

Digitonin for the Solubilization of Respiratory Complexes from the Inner Mitochondrial Membrane.

From the beginning, *n*-dodecylmaltoside has been used for the solubilization of mitochondrial respiratory protein complexes for BN-PAGE. This changed 10 years later, when the glycoside digitonin was introduced to further investigate this topic. The most striking difference between gels performed with the two detergents is the appearance of supramolecular structures in the presence of digitonin, which are clearly formed by the association of whole respiratory protein complexes to form supercomplexes. To date, digitonin is the mildest detergent available for the solubilization of membrane protein complexes. It is a natural product extracted from the plant *D. purpurea* (Purple Foxglove). Commercially available digitonin usually is offered in purities ranging from 50% to 90%. For the practical application it is important to know that the solubility of digitonin in water is limited. Heat (~ 95°C) is necessary to dissolve the detergent, which then stays in solution for only a few hours. It is somewhat surprising that digitonin is suitable to perform the solubilization of components of the inner mitochondrial membrane, because it is believed to interact exclusively with sterols. Whereas sterols are components of the outer mitochondrial membrane, the inner membrane which is housing the respiratory protein complexes, is known to lack these components almost entirely. This is probably one of the reasons why digitonin has to be used at a much higher protein-to-detergent ratio than other detergents.

The key requirement for BN-PAGE of membrane protein complexes is a solubilization process that is as mild as possible to avoid dissociation of protein complexes or supercomplexes. The goal of the solubilization process is the breakage of lipid–lipid interactions without disturbing those between proteins. Some data also suggest that even the disturbance of lipid–protein interactions can lead to the destabilization of protein complexes, because lipids might fulfill structural roles within the complex [13]. The detergents successfully employed in BN-PAGE are almost exclusively non-ionic, with only a few exceptions. The most commonly employed detergents are DDM, Triton X-100, and digitonin. To a far lesser extend, detergents like Octylglucoside, Brij96, Saponin, NP-40, and Lubrol are used. Because the different types of membranes within the eukaryotic cell also possess different lipid compositions, optimization of the solubilization method is essential for every new type of sample to be analyzed.

For soluble protein complexes, the issue of solubilization does not arise. Instead, care has to be taken to not disturb protein interactions by the use of destabilizing salt concentrations in sample, gel or cathode buffers. The stability of soluble protein complexes is affected at lower salt concentrations than that of complexes extracted from membranes. It is important to note that the presence of CBB can also cause this destabilizing effect. If a decrease in CBB concentration to a minimum does not prevent a suspected dissociation of protein complexes, it can be totally omitted from sample preparation and gel run. In this case, a so-called colorless- or clear-native

PAGE (CN-PAGE) is performed. The separation then entirely relies on the intrinsic charge of the protein complex at the given pH. However, compared to BN-PAGE, this system is limited in resolution and in the determination of the MW of the complexes [14].

Given that the protein complexes can be solubilized in a native state, their separation on BN gels depends on several factors: molecular mass (and to some extent the shape of the protein complex), physicochemical properties like hydrophobicity and intrinsic charge of the complex, as well as gel-related attributes like the slope of the acrylamide gradient in the gel, the composition of gel and running buffers, and the settings for the gel run. Because the sample-related properties have to be accepted, only the running conditions can be influenced by the experimenter. The most important parameter is the acrylamide gradient. Unlike in SDS-PAGE, where the electrophoretic mobility remains largely the same for a certain protein species throughout the entire gel lane, protein complexes appear to eventually become stuck in the BN gel. This is because during migration of the protein complexes through the gel, the acrylamide concentration rises, accompanied by a decrease in the average pore size. At a certain point, the average pore size is reduced to the degree where it is physically impossible for some complexes to migrate any further. Therefore, for a native extraction of the complexes out of the BN gel, the first dimension gel should be run only to 70–80% of the gel limit and then stopped to ensure an effective transfer of the complexes out of the first dimension gel. A variety of protocols for BN-PAGE, its second dimensions and further applications can be found in reference 15.

Second and Third Dimensions

Different second dimension gels can be combined with first dimension BN-PAGE to address different questions (Figure 38.2). Generally, these second dimensions can be divided into those with a denaturing and those with a native character. Commonly employed second dimensions are:

SDS-PAGE (Denaturing). Standard second dimension. It is probably the one best suited to provide a broad overview of the protein complexes of a given sample or for the comparison of two samples. Prior to the casting of the second dimension the BN gel strip should be incubated in a buffer containing SDS and 2-ME. This step is necessary for the complete denaturation of the protein complexes which is required for the following separation of the subunits. Preferably, the stacking gel is cast around the first dimension BN lane and uses the same gel buffer to minimize disturbances in the transition from the first to second dimension gel. A Tricine gel system consisting of a stacking, spacer, and separating gel is best for visualization and separation of samples containing proteins of less than 15 kDa. Satisfactory results can also be obtained by using a glycine-based system, especially when a better resolution of proteins larger than 15 kDa is required.

IEF-SDS-PAGE (Denaturing). This so-called 3D PAGE [16] has the advantage of a much higher resolution power compared to normal BN/SDS-PAGE. This

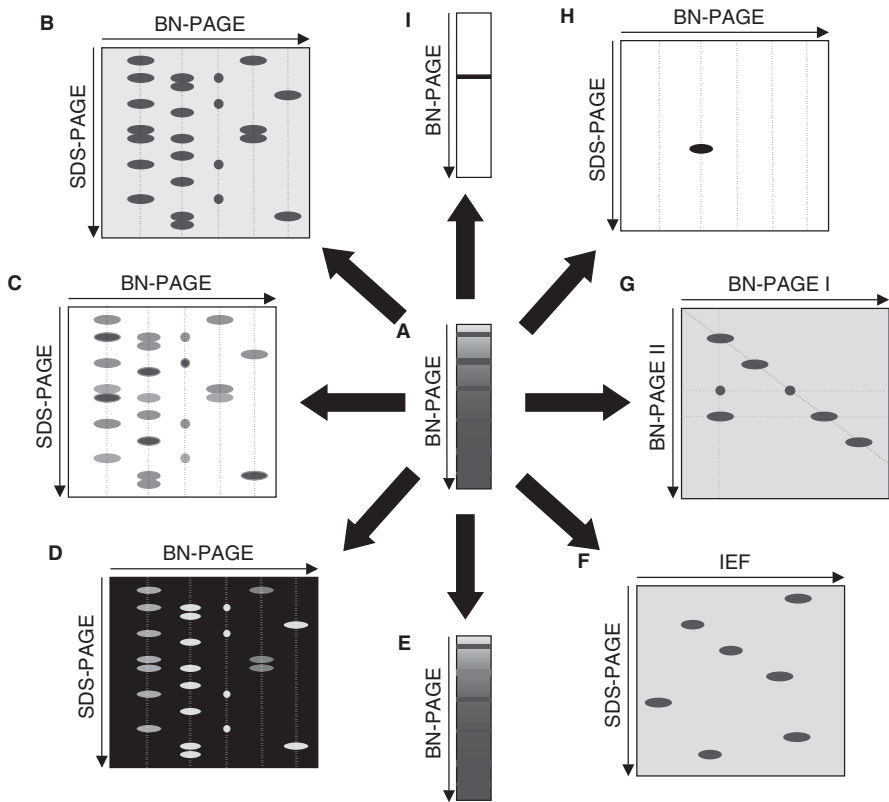


FIGURE 38.2. Schematic illustration of common downstream applications of 1D BN-PAGE. **(A)** 1D BN PAGE, unstained, **(B)** BN/SDS-PAGE for the resolution of protein complex subunits. CBB stained. **(C)** BN/SDS-PAGE for the resolution of protein complex subunits. Silver stained. **(D)** BN/SDS PAGE for the resolution of protein complex subunits. DIGE, pre-PAGE labeling of the proteins using different fluorophores for two different samples. Subunits depicted in yellow are equally abundant in both samples. Those displayed in red or green are more abundant in sample 1 or 2. **(E)** In-gel activity stain of a 1D BN-PAGE, also possible after 2D BN/BN-PAGE. **(F)** 3D BN/IEF/SDS-PAGE of a single band cut out from a 1D BN-PAGE. Proteins have been electro-eluted prior to IEF. **(G)** 2D BN/BN-PAGE. CBB stained. **(H)** Immunoblot of a 2D BN/SDS-PAGE. **(I)** Immunoblot of a 1D BN-PAGE. See insert for color representation of this figure.

is because the polypeptide subunits are resolved by two different parameters: pI and MW. It is very valuable if the protein complex consists of a high number of proteins (>30) and is also helpful for the separation of isoforms of subunits if these possess a similar MW but differ in their pIs. A drawback of this approach is the increased experimental effort required: For every single protein complex on the BN gel, a separate IEF/SDS-PAGE has to be performed. However, the introduction of weight-matched fluorophores for the labeling of different samples prior to the gel run (differential gel

electrophoresis) can reduce the experimental effort: Up to three different complexes can be resolved into their subunits on a single IEF/SDS-PAGE.

BN-PAGE (Native). This type of second dimension is suited to analyze the protein complex composition of a supercomplex [1]. The stability of protein complexes in a BN gel is dependent on the choice of detergent. Digitonin, for example, is good for the solubilization of intact supercomplexes in particular samples, while the use of DDM leads to their partial dissociation. It is therefore possible to treat a first dimension BN lane of digitonin-solubilized supercomplexes with a solution of DDM in order to dissolve these structures into their single protein complexes. After second dimension, supercomplexes and protein complexes not harmed by the harsher detergent form a diagonal line on the resulting BN/BN gels. Meanwhile, complexes or supercomplexes which were specifically destabilized under the conditions of the second gel dimension migrate below the diagonal to form vertical lines. The identity of these dissociation products can then be verified simply by comparing their MW to that of supercomplexes or protein complexes on the diagonal. This can be confirmed by MS of the complex components. As an alternative to detergent treatments, the BN strip of the first gel dimension can also be subjected to other mildly denaturing conditions such as elevated temperature, salt concentration, reductants, or other suitable chemicals. If BN/BN gels are intended to be used to perform in-gel activity stains, contact of the protein complexes with highly reactive APS and TEMED must be avoided to sustain activity of the enzymes. Therefore, it is necessary to cast the second dimension gel first and place the first dimension lane on top of it only after polymerization is completed [17].

Staining Procedures

All commonly employed staining procedures are applicable for BN-PAGE (Figure 38.2). This includes post-PAGE stains of the CBB, silver and fluorescent types, and pre-PAGE stains utilizing fluorophores for DIGE. Additionally, in-gel activity stains for the respiratory complexes I, II, IV, and V [18] as well as for POXs [19] can be undertaken after 1D BN-PAGE or 2D BN/BN-PAGE.

Other Downstream Applications of BN-PAGE

After 1D BN-PAGE, protein complexes can be blotted onto a membrane (Figure 38.2). Special precautions have to be undertaken to reduce the amount of CBB on the gels because this might later interfere with immunostaining. Usually, the amount of CBB in the cathode buffer can be reduced after one-third of the gel run time by substituting the buffer with one containing only 10% of the original CBB concentration. Electrophoresis can then be continued without any disadvantage. For some applications (e.g., 3D BN/IEF/SDS-PAGE) it might be necessary to extract a protein complex from the gel. This can happen either in a denatured state or, in some cases, in a native state, as has been demonstrated for the ATP synthase complex. After elution from the BN gel, the native electro-eluted protein complex has been stripped of the CBB dye and was subsequently used for structural analysis [20].

38.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Plant Electron Transport Chains

BN-PAGE has been used in plant science mainly to investigate the structure of the plastidic and mitochondrial electron transfer chains. Interestingly, proteins putatively functioning as carbonic anhydrases have been found to be structurally important for respiratory complex I [21]. BN-PAGE also showed that complex I of *Arabidopsis* houses the final enzyme of the pathway for the synthesis of ascorbic acid [3]. Resolution of complex III by BN/SDS-PAGE has allowed investigation of the core proteins of cytochrome reductase from potato [22] and the co-evolution of these proteins in different organisms [23]. Also, proteins responsible for the development of CMS have been identified by BN-PAGE as being subunits of complex V [24, 25]. BN-PAGE has further been employed for larger-scale investigations of the respiratory chain [16, 26], aiming to provide a broader overview of plant mitochondrial protein complexes and their subunit composition. However, complete complexes II and IV were missing on most of these gels. This changed with the introduction of digitonin for the solubilization of mitochondrial membrane protein complexes. Digitonin led to the stabilization of these complexes on the gels and facilitated the discovery of plant-specific subunits within them [2]. It also prevented the dissociation of respiratory supercomplexes. Interactions between the complexes I & III, I & III & IV, and III & IV as well as dimeric ATP synthases were found [27]. Several studies have also used BN-PAGE for the analysis of photosynthetic protein complexes and the plastidic F_0F_1 -ATP synthase [28, 29]. Again, the utilization of digitonin has facilitated the solubilization of protein supercomplexes of the plastidic electron transfer chain. Dimeric and trimeric versions of PS I have been found. Additionally, dimeric PS II participates in the formation of supercomplexes with several copies of LHC II [12]. These findings have been largely confirmed using crystallography and EM followed by single-particle analysis [for review see in reference 30]. However, in single-particle analysis of the PS I supercomplexes, resolution of the average pictures was very low, indicating either an unspecific or flexible association of the complexes [31].

Other Applications in Plants

In a number of publications, BN-PAGE has been used for the survey of the mitochondrial and plastidic protein import apparatus [32, 33] and the thylakoid protein insertion systems [34]. It has been employed for the characterization of the tobacco plastid-encoded RNA-polymerase complex [35] and plastidic ω -3 saturases [36] as well as a 350-kDa plastidic ClpP protease complex [37]. BN-PAGE has also been used for the analysis of the cytochrome c maturation complex [38], an acetyl-coenzyme A carboxylase [39], an isovaleryl coenzyme A dehydrogenase [40], the formate dehydrogenase complex [41] and several other soluble mitochondrial [42] and plastidic protein complexes. Protein complexes of the PM of spinach have been successfully resolved by BN-PAGE [43], as have complexes in the peribacteroid membrane from *L. japonicus* root nodules [44]. Also, the association of γ -tubulin complexes with microsomal membranes in *Arabidopsis* [45] has been analyzed by BN-PAGE. Many

of these results were obtained using antibodies to identify subunits of low abundant protein complexes.

38.5 CONCLUSIONS

As indicated in Section 38.4, BN-PAGE is beginning to emerge as a universal tool to study stable protein interactions on a large scale. The power of this approach in comparison to other, more focused techniques lies in its experimental simplicity and the broadness of its potential application. Within a single experiment, the subunits of several protein complexes can be separated, ready for subsequent identification. The introduction of BN-PAGE has increased our knowledge of the composition of the respiratory chain considerably. Especially in plants, but also in other organisms, a wealth of additional subunits in individual protein complexes has been found. Additionally, associations of photosynthetic and respiratory complexes into supercomplexes with defined stoichiometry have been discovered using this tool. Apart from their role in the analysis of protein interactions, 2D and 3D BN-PAGE are also proving to be a valuable tools to resolve the individual proteins present in a given membrane. BN-PAGE can act as a good compromise between IEF/SDS-PAGE, which cannot display the majority of hydrophobic proteins, and 1D SDS-PAGE, which can display hydrophobic protein but does not provide sufficient resolution of complex samples. It has been found that the average GRAVY scores of proteins derived from an *Ara-bidopsis* mitochondria are significantly higher for BN/SDS-PAGE when compared to IEF/SDS-PAGE, but these proteins also feature a wider range of hydrophobicity as indicated by the standard deviation of the GRAVY scores from the mean [46]. However, despite all the successful applications of BN-PAGE, we think that this technique still has not reached its full potential in its wider applications in proteomics and in the analysis of protein interactions in plants. This might be due to certain generic limitations of the system (see Box 38.1 and Section 38.1), but is mainly due to the fact that BN-PAGE is still not as established and widespread as IEF/SDS-PAGE in proteomic laboratories.

38.6 FIVE-YEAR VIEWPOINT

The classic starting material for BN-PAGE historically has consisted of isolated organelles that possess a relatively small amount of highly abundant protein complexes and therefore can be regarded as the ideal sample. In our opinion, the future of BN-PAGE for the investigation of other, more complex samples will strongly depend on three factors:

1. *Development and/or employment of detergents capable of stabilizing even fragile membrane protein complexes and supercomplexes:* This is especially important for the analysis of low-abundance protein complexes, because a loss of a considerable proportion of the endogenous complex during sample preparation would further reduce the abundance of the complex on the gel.

2. *Establishment of prefractionation methods compatible with BN-PAGE in order to reduce the number of different protein complexes on the gels:* Many different procedures like subfractionation of the starting material, differential solubilization, or ion exchange chromatography are putative candidates for this. Apart from avoiding disintegration of the complexes by the procedure, the main requirement is a procedure that does not prefractionate the complexes according to MW.
3. *Development of methods to closely link BN-PAGE with the structural analysis of protein complexes:* In the past, BN-PAGE, especially in the form of its 2D and 3D applications, has been a useful tool to determine the subunit composition of protein complexes, but so far it has only played a minor role in the elucidation of the structural properties of protein complexes. The development of preparative BN-PAGE in analogy to preparative SDS-PAGE has the potential to deliver purer starting material for downstream applications involved in structural analysis.

REFERENCES

1. Schägger H and Pfeiffer K. 2000. *EMBO J* **19**:1777–1783.
2. Millar AH, et al. 2004. *Plant Mol Biol* **56**:77–90.
3. Millar AH, et al. 2003. *Plant Physiol* **133**:443–447.
4. Kleffmann T, et al. 2004. *Curr Biol* **14**:354–362.
5. Millar AH, et al. 2001. *Plant Physiol* **127**:1711–1727.
6. Perales M, et al. 2004. *Plant Mol Biol* **56**:947–957.
7. Peltier JB, et al. 2006. *Mol Cellular Proteomics* **5**:114–133.
8. Heazlewood JL, et al. 2006. *Nucleic Acids Res* **35**:D213–D218.
9. Schägger H and von Jagow G. 1991. *Anal Biochem* **199**:223–231.
10. Huang D, et al. 1994. *Biochemistry* **33**:4401–4409.
11. Eubel H, et al. 2004. *Plant Physiol Biochem* **42**:937–942.
12. Heinemeyer J, et al. 2004. *Phytochemistry* **65**:1683–1692.
13. Pfeiffer K, et al. 2003. *J Biol Chem* **278**:52873–52880.
14. Wittig I and Schägger H. 2005. *Proteomics* **5**:4338–4346.
15. Wittig I, Braun HP, and Schägger H. 2006. *Nature Protocols* **1**:418–428.
16. Werhahn W and Braun HP. 2002. *Electrophoresis* **23**:640–646.
17. Eubel H, Heinemeyer J, and Braun HP. 2004. *Plant Physiol* **134**:1450–1459.
18. Sabar M, Balk J, and Leaver CJ. 2005. *Plant J* **44**:893–901.
19. Fecht-Christoffers MM, et al. 2003. *Plant Physiol* **133**:1935–1946.
20. Poetsch A, et al. 2000. *Biochim Biophys Acta—Biomembranes* **1466**:339–349.
21. Perales M, et al. 2005. *J Mol Biol* **350**:263–277.
22. Jansch L, et al. 1995. *Eur J Biochem* **228**:878–885.
23. Brumme S, et al. 1998. *J Biol Chem* **273**:13143–13149.
24. Heazlewood JL, Whelan J, and Millar AH. 2003. *FEBS Lett* **540**:201–205.

25. Sabar M, et al. 2003. *EMBO Rep* **4**:381–386.
26. Heazlewood JL, et al. 2003. *Plant Physiol* **132**:230–242.
27. Dudkina NV, et al. 2006. *Trends Plant Sci* **11**:232–240.
28. Ciambella C, et al. 2005. *Proteomics* **5**:746–757.
29. Rexroth S, et al. 2004. *Biochim Biophys Acta—Bioenergetics* **1658**:202–211.
30. Dekker JP and Boekema EJ. 2005. *Biochim et Biophys Acta—Bioenergetics* **1706**:12–39.
31. Kouril R, et al. 2005. *Photochem Photobiol Sci* **4**:1091–1094.
32. Kikuchi S, Hirohashi T, and Nakai M. 2003. *Plant Cell Physiol* **44**:S203–S203.
33. Werhahn W, Jansch L, and Braun HP. 2003. *Plant Physiol Biochem* **41**:407–416.
34. Klostermann E, et al. 2002. *Biochem J* **368**:777–781.
35. Suzuki JY, et al. 2004. *Plant J* **40**:164–172.
36. Sakamoto H, et al. 2004. *Plant Cell Physiol* **45**:S116–S116.
37. Peltier JB, et al. 2001. *J Biol Chem* **276**:16318–16327.
38. Giege P, et al. 2004. *FEBS Lett* **563**:165–169.
39. Focke M, et al. 2003. *Plant Physiol* **133**:875–884.
40. Daschner K, Couee I, and Binder S. 2001. *Plant Physiol* **126**:601–612.
41. Bykova NV, et al. 2003. *J Biol Chem* **278**:26021–26030.
42. Giege P, Sweetlove LJ, and Leaver CJ. 2003. *Plant Mol Biol Rep* **21**:133–144.
43. Kjell J, et al. 2004. *Physiol Plant* **121**:546–555.
44. Wienkoop S and Saalbach G. 2003. *Plant Physiol* **131**:1080–1090.
45. Drykova D, et al. 2003. *Plant Cell* **15**:465–480.
46. Eubel H, Braun HP, and Millar AH. 2005. *Plant Methods* **1**:11.
47. Witting I, Schagger H. 2005. *Proteomics* **5**:4338–4346.

PROTEIN–PROTEIN INTERACTION MAPPING IN PLANTS

Joachim F. Uhrig

39.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Proteins operate mainly through interactions with other molecules such as nucleic acids, lipids, or LMW compounds. The monomeric form of a protein is hardly ever fully functional; instead, the formation of transient or stable protein complexes is principally a necessary prerequisite for biological activity. Consequently, PPIs are fundamental to virtually every aspect of biological functions. With the aim of understanding cellular processes on a systems level, comprehensive protein-linkage maps or “interactome” analyses are one of the major goals in post-genomic research [1]. Facilitated by the development of high-throughput technologies allowing the mass analysis of protein complexes and PPIs, huge amounts of data have accumulated and are now beginning to reveal highly complex and intricate networks of interactions. To meet the challenge of interpreting these networks, sophisticated mathematical tools have been developed; and in the past few years, considerable advances in the graph-theoretical analysis of complex networks have been achieved [2, 3]. General principles underlying the architecture of networks as diverse as the World Wide Web, co-authorships on scientific publications, or the pandemic spread of infectious diseases have been

revealed, and the notion that this universal architecture can be attributed to biological networks as well has resulted in much attention and excitement [4–6]. The modular structure of complex networks apparently reflects the modular nature of biological organization, and the association of cellular functions with quantifiable patterns in network topology is increasingly regarded as a promising novel way to achieve a deeper understanding of complex biological processes [7].

Biological functions are rarely accomplished by the action of single molecules, but similarly usually do not involve the interplay of thousands of different cellular components. To describe cellular processes beyond the single molecule level, Hartwell and co-workers introduced the concept of “functional modules.” This idea proposes the existence of isolated, functional units at the so-called meso-scale level (i.e., five to 25 molecular components) acting autonomously [8, 9]. Such functional modules that are required to be chemically or spatially separable from the whole system have been characterized experimentally *in vitro* and *in vivo* [8]. The global topology of complex molecular networks is envisioned to provide insight into large-scale organizational principles of biological systems. Functional modules, on the other hand, may have more direct relevance for understanding discrete cellular functions. In fact, functional modules emergent in molecular interaction networks are increasingly recognized as the operational units of biomolecular systems [8–13]. Particularly with respect to protein interaction networks, a great deal of interest is currently focusing on smaller-scale topological structures, the modules, motifs, and cliques (see Box 39.1 for details on graph theoretical terms). In contrast to conclusions from global network properties, information and hypotheses derived from local topologies of protein interaction networks may be more easily amenable to experimental validation and can be instrumental in designing novel experimental strategies such as the “interactome walking” (see below) that recently has been successfully applied to unravel the TGF- β signaling pathway in *C. elegans*, the DNA replication network of *B. subtilis*, and a complex protein interaction network associated with Huntington Disease [14–16].

On the structural level, protein interactions are based on accurately complementary shapes of protein interfaces with characteristic distribution of charged, polar, and hydrophobic amino acids. In-depth analyses of protein interaction surfaces revealed that usually only a few strongly conserved residues, so-called “hot-spots,” contribute cooperatively to the strength and the specificity of PPIs [17, 18]. Evolutionary constraints result from the fact that a mutation affecting one interaction surface has to be compensated for by a corresponding mutation on the other interaction surface. However, gene duplication leading to the presence of two gene products with the same interaction properties, instantly releases the evolutionary constraints. Gene duplication and subsequent functional divergence is one of the fundamental mechanisms of the evolution of genomes and complexity in general; and interestingly, the same mechanism shapes the evolution of complex molecular networks as well [19, 20]. Protein interactions can be regarded one of very few general measures of protein functions that can be analyzed systematically and on a genome-wide scale [21]. In the yeast interactome network, duplicated genes share only very few interactions; in other words, interactions diverge very rapidly, indicating in turn that the divergence of protein functions after gene duplication is, in evolutionary terms, a very fast process

Box 39.1 Basic Concepts of Graph-Theoretical Network Analysis

Networks are graphical visualizations of elements termed **nodes**, connected by **links** indicating their relationship. In biological networks, nodes can be, for example, genes or proteins and links can indicate co-expression or interaction. Properties of the entire network, as well as characteristics of single nodes, single links, or subgraphs, can be described and quantified mathematically. The most basic attribute of a node in a network is its **degree k** , which is defined as the number of links connecting it with other nodes, and an elementary measure to characterize a whole network's topology is the **degree distribution $P(k)$** , obtained by counting the numbers of nodes having the same degree $N(k)$ divided by the total number of nodes (N). $P(k)$ gives the probability of a node having exactly the degree k . An exciting discovery of recent years, pointing at universal principles behind complex evolving networks, is that many technical networks, social networks, and various biological networks share a very similar degree distribution following a **power-law ($P(k) \sim k^{-\gamma}$)** with a **degree exponent γ** between 2 and 3. Another basic characteristic used to describe general network architecture and the relationships between nodes is the **path length**, which is defined as the number of steps needed to get from one node to another. The **shortest path** and the **mean path length**, defined as the average over the shortest paths of every node to every other node, are of special interest and are measures of the diameter of a network. Complex networks have very short mean path lengths, a property characteristic of random networks. On the other hand, biological and other complex networks exhibit a high degree of clustering that is not found in random networks but rather is an attribute of regular networks. In many networks it has been found that if, for example, a node A is connected to B and C, it is highly probable that B has a direct link to C, or, in other words, that A, B, and C form a triangle. This observation can be quantified using the **clustering coefficient C_i** , defined as the number of links (n) between the neighbors (k) of a node i divided by the maximum number of links possible between these neighbors: $C_i = 2n_i / k(k-1)$. A high clustering coefficient is indicative of a "**community structure**" and a modular organization of the network. **Modules** emerging from network representations are assemblies of nodes more closely interconnected amongst each other than to the rest of the network. Functional annotation of these separable subgraphs in biological networks indicates that these structures reflect the modularity of cellular functions. The elementary building blocks of **functional modules** are small patterns termed **network motifs** that recur at frequencies significantly higher than those found in equivalent randomized networks. These motifs are simple geometrical figures like triangles, squares, or pentagons with a certain degree of internal connections. Completely connected subgraphs/motifs—that is, geometrical structures in which every node is linked to every other node—are called **cliques**. Significant overrepresentation of certain motifs or cliques has been shown to be of functional relevance and might be used to functionally distinguish different types of networks.

[19–21]. Thus, systematic analyses of protein interaction specificities within protein families—the products of multiple gene duplications—may be a straightforward way to dissect functional redundancy/specificity to understand structure–function relationships. This general idea is underlying a number of recent medium-scale projects in plant research described in Section 39.3.

In comparative genomics, the question of functional specificity is becoming increasingly important. The transfer of knowledge between organisms depends on unambiguous identification of orthologous genes or proteins. Orthologs, however, cannot simply be inferred from sequence alone, and surveys taking shared interaction partners as a measure of functional overlap recently revealed that experimentally identified functional classes are not identical to groups clustered on the basis of sequence similarity [23–26]. Experimentally derived systematic information on interaction specificities can be instrumental in revealing structural rules beyond sequence homology. Such detailed knowledge in turn can be useful for a more accurate prediction of functional orthologs between species [25].

To date, partial interactomes for a number of eukaryotic model organisms including yeast, *C. elegans*, and *D. melanogaster* are available, and recently the first steps toward a comprehensive human protein interaction map have been reported [27–33]. In plants, however, protein interaction mapping is lagging considerably behind. Even though a few large-scale projects have been initiated—aiming, for example, at a complete picture of TF interactions [34, 35] or the membrane protein interactome of *A. thaliana* (<http://www.associomics.org>)—comprehensive experimental data sets are not available at present. All the same, even with these data still missing, advance in the graph theoretical functional interpretation of complex network architectures might pave the way for novel approaches in plant research. Exemplified by several recently published projects, particularly the intermediate and local levels of network organization—the modules, motifs, and cliques—promise to be a valuable tool for designing focused experimental approaches aiming at an understanding of functional specificities and novel regulatory connections and pathways [36].

39.2 METHODOLOGY AND STRATEGY

A number of technologies have been developed to detect and monitor PPIs *in vitro* or *in vivo* [37]. However, only few have been proven sufficiently robust and adaptable to be used in large-scale protein interaction mapping projects. By far the majority of available protein interaction data have been obtained with the Y2H system [38]. As a complementary approach, affinity purification of tagged protein complexes in conjunction with MS/MS has been developed as a high-throughput technology [39]. The latter technology has been applied on a genome-wide or near-genome-wide scale to yeast and *E. coli*, so far [40–42]. TAP-tagging in plants followed by affinity purification and MS analysis of protein complex composition is described elsewhere in this volume (see Chapters 36 and 37). This chapter, therefore, will focus on the concepts and strategies of large-scale Y2H experiments for systematic protein interaction mapping.

The Y2H system has been applied to a broad spectrum of different proteins, and a number of variants and different methodological and technological strategies

have been developed to facilitate large-scale or high-throughput application of this method. Generally, the two proteins to be examined for interaction are co-expressed in yeast as fusion proteins. The so-called “bait” protein, on the one hand, is fused to a DNA-binding domain from a TF, the “prey” protein, on the other hand, is fused to a transcription activation domain. When there is an interaction between bait and prey, reporter genes are activated that can be monitored by allowing prototrophic growth on selective media or by measuring enzymatic activities.

Scaling up the Y2H method requires dealing with large numbers of transformed yeast cells bearing two plasmids encoding the two fusion proteins to be assayed. Despite the development of high-efficiency transformation protocols, transforming yeast cells with two plasmids or with complex cDNA libraries is rather laborious and not altogether very efficient. A far more efficient way of obtaining yeast cells expressing fusion proteins from two plasmids is the so-called interaction mating. Interaction mating makes use of the yeast life cycle with its haploid and diploid phases. If the nutritional status is appropriate, haploid cells of opposite mating type (MATa and MAT α) can conjugate and form diploid zygotes. Under the right conditions, this is a rather efficient process and can be used to combine two plasmids transformed individually into the same diploid zygote. Moreover, in contrast to a chemical transformation procedure that inevitably poses very much stress to the cell, zygotes are very fit and viable, resulting in faster growth and therefore shorter incubation times for the positive colonies to develop.

Three strategies are currently pursued to perform systematic Y2H projects: First, the direct one-by-one testing of every combination of proteins in Y2H matrices or arrays, second, the utilization of defined pools of bait and prey constructs; and third, the identification of interacting protein pairs by screening genomic or cDNA libraries (Figure 39.1).

The most straightforward way to obtain a genome-wide protein linkage map, obviously, is to test every single protein directly for interaction with every other protein. In a comprehensive fashion, a number of organisms with small genomes (e.g., bacteriophage T7 and, partially, the yeast genome) have been analyzed in this way [27, 43]. However, there are technical limitations that prevent this matrix approach from being applied to larger genomes: The combinatorics of an all-against-all assay of large clone sets by far exceeds the capacity even of advanced picking and spotting robots or automated technologies, allowing high-throughput liquid handling.

Therefore, pooling strategies allowing the simultaneous testing of large numbers of clones have been developed and successfully applied in high-throughput Y2H projects. Aiming at comprehensive protein linkage maps of multicellular eukaryotic model organisms, researchers have published comprehensive protein linkage maps of *D. melanogaster* and *C. elegans*, and, recently, the first draft versions of a partial human interactome. In one approach toward the human interactome, Stelzl et al. [33] assayed 557 pools of eight bait proteins each for interaction with 5632 individual prey proteins. On the other hand, Rual et al. [32] tested the pairwise interactions amongst 8100 human proteins by mating single baits to pools containing 188 preys each. In total, these two projects identified more than 6000 PPIs, putting a number of human disease-associated proteins into a broader functional context.

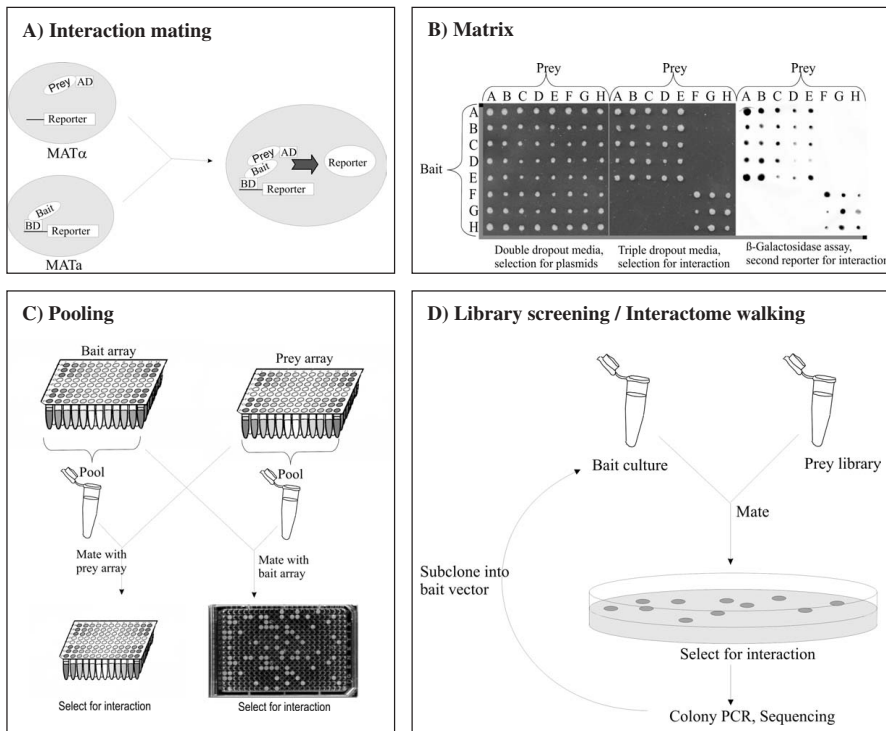


FIGURE 39.1. Y2H high-throughput screening strategies. **(A)** Yeast mating is generally used in all large-scale Y2H approaches to bring together bait and prey plasmid in the diploid zygotes. **(B)** In the matrix approach, haploid strains harboring single bait or prey plasmids, respectively, are mated and tested individually. **(C)** Pooling allows the simultaneous testing of multiple baits or preys, respectively. More sophisticated pooling schemes have been developed allowing further increasing the throughput of the screening. **(D)** Screening libraries with the Y2H system allows the unbiased identification of candidate interactors that in turn can be used as baits in subsequent screenings to build up an interaction network.

The matrix approach and the analysis of defined pools of baits and preys rely on the availability of comprehensive collections of cloned full-length coding sequences of every gene of a genome, so-called ORFeomes. Higher eukaryotes have predominantly split genes with exon and intron regions, precluding the generation of ORFeomes by simple amplification of ORFs directly from genomic DNA, which has been possible in the case of procaryotes and lower eukaryotes such as the yeast *S. cerevisiae*. Therefore, more laborious experimental strategies are necessary to assemble ORFeomes, making this a considerably difficult task. However, implementation of recombination cloning systems based upon lambda attachment sites or the cre/lox system are increasingly used to standardize cloning of cDNAs or genomic fragments into bait or prey vectors, respectively, and, furthermore, can greatly simplify the construction of Y2H libraries [44]. ORFeomes are not available for plants, so far, although ORFeome projects have

been initiated for *A. thaliana*, and it can be expected that a full set of expression clones will be available in the near future(<http://www.evry.inra.fr/public/projects/orfeome/orfeome.html>).

As a third strategy, the screening of libraries allows the identification of PPIs on a large scale. Here, single bait constructs are co-expressed in yeast with libraries of genomic fragments or cDNA, respectively, ligated into prey vectors. Colony growth on media selecting for activity of the reporter genes indicates the presence of a prey protein interacting with the bait protein. While a positive signal in Y2H matrix experiment already reveals which bait protein interacts with which prey protein, a positive colony from a library screening has to be analyzed further by isolating the candidate prey clone and identifying its identity by sequencing.

The particular functional importance of local topologies—modules, motifs, and cliques—in protein interaction networks is the rationale of the so-called “interactome walking” approach that focuses on the direct neighbourhood of particular bait proteins or protein families, allowing to successively build up a local network of potential functional relevance. In interactome walking, prey proteins identified from an initial library screening are subcloned into bait vectors and used as baits in secondary and tertiary rounds of screenings. Except for the limitations posed by the quality and the comprehensiveness of the libraries used, this is a completely unbiased approach allowing the identification of novel functional links [1].

Screening a complex cDNA library is a rather big effort because large numbers of double transformants—usually several millions—are needed to warrant complete coverage. Recent technological improvements have, however, greatly simplified the interaction mating procedure, in a way that multiparallel library screenings on large-scale are possible and have been successfully applied to several bacteria and *S. cerevisiae*; also in plant research, the first successful application of systematic library screenings to obtain protein interaction maps have been reported [45–48].

While the importance of large-scale protein interaction data as a valuable source of functional information is undisputed, a critical issue regarding such high-throughput data sets is the quality of the data. Low overlap in the results of the two genome-wide Y2H projects has raised concerns about FP or FN results [27, 49].

The Y2H system is an *in vivo* method requiring the protein interaction to take place in the yeast nucleus. This excludes the analysis of interactions depending on specific PTMs not available in yeast, or relying on particular intracellular localization such as interactions of membrane proteins. Moreover, an intrinsic limitation of the Y2H system is that proteins bearing transcriptional activating potential cannot be utilized as bait proteins. When using single or small numbers of bait proteins, every protein has to be tested individually for prey-independent activation of the reporter gene. This, however, may be rather laborious when using large numbers, or even libraries of bait proteins. However, with the development of a reverse two-hybrid system a general pre-selection scheme selecting against such auto-activating properties of bait proteins is available to prepare libraries of bait proteins suitable for screening.

Despite these limitations and drawbacks, extensive efforts have been taken to assess the relevance and the reliability of available high-throughput Y2H data and have proven a reasonably good quality [50]. The major shortcoming of these large-scale

projects seems to be their intrinsic high FN rate. Focused smaller-scale projects might in this respect be more efficient in detecting a maximum amount of reliable interaction pairs [31, 51].

39.3 EXPERIMENTAL RESULTS AND APPLICATIONS

While at present, comprehensive interactome data are not available for any plant species, a number of medium-scale projects have proven the value of systematic protein interaction mapping approaches to analyze and understand regulatory networks in plants. Examples come from a number of different areas of plant research such as the regulation of cell shape and organ development, hormone and calcium signaling, and networks involved in cell cycle and plant stress responses. Most of the data have been obtained with the model organism *A. thaliana*; nonetheless, some systematic protein interaction data are available from other plant species such as rice, wheat, or petunia. Selective protein interactions and combinatorial diversity of protein complex formation are particularly important in the control of transcription [52], and several protein families of plant TFs including the TALE homeodomain protein family, the MADS box TFs, MYB proteins and bZIP TFs have been systematically analyzed with respect to their protein interaction properties [25, 26, 46, 53, 54].

MADS box TFs have a central regulatory function in a number of vegetative or reproductive developmental processes in plants. Their specificity, activity, and regulation rely to a great extent on their potential to form homo- and heterodimers as well as ternary or higher-order complexes [55]. The MIKC class MADS box proteins from *Petunia* have been analyzed with respect to their interaction specificities, and the interaction networks have been integrated with phylogenetic and expression data [53]. More comprehensively, more than 100 MADS box TFs from *A. thaliana* have been investigated for homo- and heterodimerization using the Y2H system applying a matrix approach [26]. The resulting intricate network of interactions revealed functional relationships different from those predicted from clustering according to sequence homologies. Novel regulatory loops were detected, providing unexpected links between flower organ development and floral induction [26]. The most highly connected core of the network is formed by a clique of six MADS proteins. According to the conclusions drawn from graph theoretical analyses of the yeast interactome, such cliques often represent multi-subunit protein complexes. Taking into account that functional specificity of MADS proteins is determined on the level of ternary or quaternary complexes [56], the network topology indicates that MADS proteins from different classes might achieve novel regulatory specificities by cooperating in higher order protein complexes.

Combining multiparallel Y2H library screening with an all-against-all interaction matrix approach, the TALE homeodomain protein family from *A. thaliana* has been investigated systematically [46]. This TF family, comprising the BELL and KNOX subfamilies, is involved in plant meristem function and leaf development. The resulting highly connected network consists of more than 200 interactions amongst 49 proteins including nine members of the previously uncharacterized OVATE protein

family (AtOFPs). Quantification of the connectivity and the local network patterns using graph theoretical tools revealed properties comparable to network modules of functional relevance found in the yeast interactome network. In fact, only six modules can be found in the yeast interactome network exhibiting a similar high density of connections amongst a comparable number of proteins. All of them represent either known protein complexes or components of functionally homogeneous systems [46]. The BELL and KNOX proteins have split early in plant evolution [57]. The finding that several OVATE proteins are able to interact with members of both subfamilies, therefore, indicates an evolutionary ancient functional link between TALE proteins and OVATE proteins, conserved despite the generally very fast protein interaction turnover in the evolution of protein families [22]. Implications from the network topology have been investigated experimentally. Overexpression of AtOFP1 caused severe pleiotropic developmental aberrations, and several AtOFPs were shown to regulate the subcellular localization of TALE homeodomain proteins by providing a physical link to microtubules, supporting the conclusion of OVATE proteins being a novel family of fundamental regulators of plant development [46].

Systematic interaction mapping can be used to address the question of the structural basis determining the specificity of protein interactions. In this respect, the MYB TF family from *A. thaliana* has been investigated systematically [25]. Comprising more than 130 members, MYB proteins, are one of the largest TF families in plants [58]. They cooperate with BHLH proteins in regulatory networks controlling phenylpropanoid biosynthetic pathways and epidermal cell differentiation and patterning [59, 60]. Correlating amino acid variations in the conserved MYB domain with protein interaction specificities revealed a conserved six-amino-acid signature as the structural basis of interactions between MYB and R/B-like BHLH proteins. This pattern could be used to functionally discriminate very closely related MYB proteins and to predict novel MYB/BHLH interactions from genomic sequences [25]. MYB proteins possessing this amino acid signature are phylogenetically quite divergent. Functional classification on the basis of general sequence homologies would therefore not have been possible. The identified amino acid signature is very specific and it is conserved amongst higher plants. It might therefore provide the basis for knowledge transfer from the model species *A. thaliana* to plants of agricultural importance.

Several further protein interaction mapping projects focusing on plant-specific protein families and signaling processes have been published recently. It is increasingly appreciated that selectivity of protein interactions is an important means to impose specificity to signaling pathways. Calcium signaling, MAPK cascades, and two-component systems, as well as the targeted protein degradation pathway, in plants have been systematically analyzed in this respect. Protein-specific differences in the selectivity of complex formation have been detected and have been used to functionally classify members of protein families, providing conceptual models of the mechanisms generating signaling specificity [61–63]. Regulatory specificity within the cytokinin-signaling pathway in *Arabidopsis*, for example, has been investigated by a systematic Y2H analysis of protein interactions between 31 two-component system proteins (phosphotransmitters and response regulators) acting downstream of the

cytokinin receptors. The resulting interaction network including 68 protein interactions revealed both redundancy and specificity within the system [64].

With the aim of dissecting the molecular basis of plant cell morphogenesis, a network of protein interactions centering on regulators of cytoskeleton dynamics has recently been investigated [65]. A systematic all-against-all Y2H interaction matrix comprising 31 proteins known or predicted to be involved in ARP2/3 complex-mediated actin branching revealed 68 mostly novel PPIs, bringing previously uncharacterized components such as the small Rho-like GTPase ROP7 and the putative ROP guanine exchange factor SPIKE1 into the functional context of actin-dependent regulation of cell morphogenesis [65].

As a first example of an application of the “interactome walking” strategy to plants, Langridge and colleagues identified wheat proteins involved in pre-mRNA splicing and mRNA export [47]. This project not only demonstrates the feasibility of “interactome walking” to put proteins into a functional context, but also is an example for successful knowledge transfer from the model organism *A. thaliana* to economically more important plant species like in this case wheat. A wheat homologue of the *Arabidopsis* splicing factor AtRSZ33 was used as starting bait in Y2H screenings to identify factors involved in pre-mRNA processing. Prey cDNAs identified in this screening were then subcloned into bait vectors and were used as baits in successive Y2H screenings to build up a network of interactions.

PPI data alone may not be sufficient to assign functions to unknown proteins. However, interaction networks provide a framework to integrate biological information from other approaches. Combining data from different functional genomics approaches such as proteomics data, expression profiles, systematic localization studies, mutant analyses, and metabolic profiles promises to very effectively allow the functional annotation of uncharacterized proteins. One of few examples for such an integrated approach in plants focuses on the molecular basis of developmental and stress responses in rice. Using expression profiling, mutant analysis, and a complex network of PPIs derived from systematic interaction mapping by a large-scale Y2H library screening approach, several novel proteins have been identified playing a role in disease resistance [48].

39.4 CONCLUSIONS

In recent years, significant technological developments have been achieved facilitating high-throughput protein interaction mapping that resulted in comprehensive protein interaction networks for a number of model organisms. Now eventually, medium- and large-scale protein interaction mapping projects are beginning to be established in plant research, and first results prove their value in revealing novel regulatory pathways and in dissecting the functions of uncharacterized proteins and protein families. Despite its inherent limitations and drawbacks, the Y2H system is the major source of comprehensive protein interaction data. With the accumulation of high-throughput data sets, the reliability of the data is increasingly becoming an issue. Here, knowledge from graph-theoretical analyses of interactome networks can be instrumental in revealing

reliable and functionally significant patterns (functional modules, motifs or cliques) to direct focused experimental approaches aiming at a systems-level understanding of cellular functions.

39.5 FIVE-YEAR VIEWPOINT

Sequencing of plant genomes is proceeding rapidly; and in the near future, genome sequences of a number of plant species will be available. While this will constitute a wealth of information, gaining deeper understanding of biological processes will require integrating functional genomics, proteomics, and interactomics data. For protein expression at the proteome scale, the systematic and comprehensive cloning of full-length ORFs in a standardized way will be required. Such ORFeome projects are initiated, and these resources will greatly enhance the possibilities in large-scale plant protein interaction mapping. Recent data suggest that interactome structures can be conserved between species and can therefore be instrumental in knowledge transfer and annotating protein functions across species. The availability of interactome data for several plant species will pave the way for such comparative approaches. High-throughput protein interaction data obtained with the Y2H system or likewise with MS approaches do not provide much information on the question of where and when proteins interact *in vivo*. The integration of interactome data with information from different resources such as expression profiles and phenotypic data have proven to effectively complement one another in predicting cellular functions. However, new concepts and most likely new technologies will be needed to proceed from the static picture of networks to the highly dynamic situation *in vivo*.

REFERENCES

1. Cusick ME, et al. 2005. *Hum Mol Genet* **14**:R171–R181.
2. Strogatz SH. 2001. *Nature* **410**:268–276.
3. Watts DJ and Strogatz SH. 1998. *Nature* **393**:440–442.
4. Xia Y, et al. 2004. *Annu Rev Biochem* **73**:1051–1087.
5. Alm E and Arkin fmAP. 2003. *Curr Opin Struct Biol* **13**:193–202.
6. Bray D. 2003. *Science* **301**:1864–1865.
7. Barabasi AL and Oltvai ZN. 2004. *Nat Rev Genet* **5**:101–113.
8. Hartwell LH, et al. 1999. *Nature* **402**:C47–C52.
9. Spirin V and Mirny LA. 2003. *Proc Natl Acad Sci USA* **100**:12123–12128.
10. Ihmels J, et al. 2002. *Nat Genet* **31**:370–377.
11. Rives AW and Galitski T. 2003. *Proc Natl Acad Sci USA* **100**:1128–1233.
12. Wuchty S, Oltvai ZN, and Barabasi AL. 2003. *Nat Genet* **35**:176–179.
13. Snel B and Huynen MA. 2004. *Genome Res* **14**:391–397.
14. Noiroi-Gros MF, et al. 2002. *Proc Natl Acad Sci USA* **99**:8342–8347.
15. Tewari M, et al. 2004. *Mol Cell* **13**:469–482.

16. Goehler H, et al. 2004. *Mol Cell* **15**:853–865.
17. Halperin I, Wolfson H, and Nussinov R. 2004. *Structure* **12**:1027–1038.
18. Keskin O, Ma B, and Nussinov R. 2005. *J Mol Biol* **345**:1281–1294.
19. Wagner A. 2003. *Proc Biol Sci* **270**:457–466.
20. Teichmann SA and Babu MM. 2004. *Nat Genet* **36**:492–496.
21. Wagner A. 2001. *Mol Biol Evol* **18**:1283–1292.
22. Berg J, Lassig M, and Wagner A. 2004. *BMC Evol Biol* **4**: 51.
23. Theissen G. 2002. *Nature* **415**:741.
24. Vinson C, et al. 2002. *Mol Cell Biol* **22**:6321–6335.
25. Zimmermann IM, et al. 2004. *Plant J* **40**:22–34.
26. de Folter S, et al. 2005. *Plant Cell* **17**:1424–1433.
27. Uetz P, et al. 2000. *Nature* **403**:623–627.
28. Ito T, et al. 2001. *Proc Natl Acad Sci USA* **98**:4569–4574.
29. Giot LB, et al. 2003. *Science* **302**:1727–1736.
30. Li S, et al. 2004. *Science* **303**:540–543.
31. Formstecher E, et al. 2005. *Genome Res* **15**:376–384.
32. Rual JF, et al. 2005. *Nature* **437**:1173–1178.
33. Stelzl U, et al. 2005. *Cell* **122**:957–968.
34. Paz-Ares J and The REGIA Consortium. 2002. *Comp Funct Genomics* **3**:102–108.
35. Kersten B, et al. 2002. *Plant Mol Biol* **48**:133–141.
36. Uhrig JF. 2006. *Planta* **224**:771–781.
37. Piehler J. 2005. *Curr Opin Struct Biol* **15**:4–14.
38. Fields S. 2005. *Febs J* **272**:5391–5399.
39. Dziembowski A and Seraphin B. 2004. *FEBS Lett* **556**:1–6.
40. Gavin AC, et al. 2002. *Nature* **415**:141–147.
41. Butland G, et al. 2005. *Nature* **433**:531–537.
42. Ho Y, et al. 2002. *Nature* **415**:180–183.
43. Bartel PL, et al. 1996. *Nat Genet* **12**:72–77.
44. Walhout AJ, et al. 2000. *Methods Enzymol* **328**:575–592.
45. Soellick TR and Uhrig JF. 2001. *Genome Biol* **2**:52–57.
46. Hackbusch J, et al. 2005. *Proc Natl Acad Sci USA* **102**:4908–4912.
47. Lopato S, et al. 2006. *Plant Mol Biol* **62**:637–653.
48. Cooper B, et al. 2003. *Proc Natl Acad Sci USA* **100**:4945–5490.
49. Ito T, et al. 2000. *Proc Natl Acad Sci USA* **97**:1143–1147.
50. Bader JS, et al. 2004. *Nat Biotechnol* **22**:78–85.
51. Stanyon CA, et al. 2004. *Genome Biol* **5**:R96.
52. Wolberger C. 1999. *Annu Rev Biophys Biomol Struct* **28**:29–56.
53. Immink RG, et al. 2003. *Mol Genet Genomics* **268**:598–606.
54. Ehlert A, et al. 2006. *Plant J* **46**:890–900.
55. Messenguy F and Dubois E. 2003. *Gene* **316**:1–21.
56. Kaufmann K, Melzer R, and Theissen G. 2005. *Gene* **347**:183–198.
57. Bellaoui M, et al. 2001. *Plant Cell* **13**:2455–2470.

58. Stracke R, Werber M, and Weisshaar B. 2001. *Curr Opin Plant Biol* **4**:447–456.
59. Broun P. 2005. *Curr Opin Plant Biol* **8**:272–279.
60. Zhang F, et al. 2003. *Development* **130**:4859–4869.
61. Muller J, et al. 2001. *Plant J* **27**:13–23.
62. Batistic O and Kudla J. 2004. *Planta* **219**:915–924.
63. Risseuw EP, et al. 2003. *Plant J* **34**:753–767.
64. Dortay H, et al. 2006. *Febs J* **273**:4631–4644.
65. Uhrig JF, et al. 2007. *Development* **134**:967–977.

PART VII

PLANT DEFENSE AND STRESS

PROTEOMICS IN PLANT DEFENSE RESPONSE

Sun Tae Kim and Kyu Young Kang

40.1 INTRODUCTION

The innate immune system of plants is responsible for disease resistance through recognition, directly or indirectly, of proteins or other compounds that are secreted by invading microbes called “avirulence” factors or elicitors. The initial recognition event is mediated by direct or indirect interactions between the pathogen’s avirulence factor and the corresponding host resistance protein—that is, the gene-for-gene model. The biochemical changes within plant invaded by pathogen are very important because they may give insights to understand how defense mechanisms during plant–microbe interactions (PMIs) could be controlled.

In recent years, advanced molecular approaches have been used for the identification of genes and functions involved in PMIs. Studies of gene expression involved in PMI at the transcriptional level have been led by differential screening methods like differential display, RT-PCR, suppressive subtractive hybridization, SAGE, and DNA chip. Obviously, transcriptomics data using DNA chip have been comprehensively used not only to classify differential regulation and expression of genes in response to pathogens, but also to identify known and unknown defense genes. Although a

number of researchers have studied on analyzing gene expression on the transcript level, however, mRNA-based approaches have some limitations. First, mRNA levels do not necessarily correlate with protein levels. Second, mRNA analyses give us little information about whether the encoded proteins are active or not; and the functions of proteins are often regulated by PTMs such as phosphorylation and glycosylation, processes that can be vital for cellular functions. Third, mRNA cannot be used to profile changes in secreted proteins when the plant leaves are invaded by pathogens. Finally, PPIs and the molecular composition of cellular structures can be determined only at the protein level. Thus, an important complementary approach for investigating plant defense mechanism can be provided by proteomics studies.

With the completion of genome sequencing projects and the steady advancements in the tools such as 2-DGE, mass spectrometer, MALDI-TOF, and ESI-MS/MS for protein identification and characterization, proteomics provides new opportunities for studying complex processes at the protein level. This technology allows us to investigate not only complete proteomes at the organelle, cell, organ, or tissue levels, but also differential proteomes affected by different physiological conditions such as pathogen. This is also a meaningful measure for identifying a set of proteins that can be tested for functional involvement in plant pathogen immune responses. Figure 40.1 shows a proteomics approach for defense proteome analysis. Many comprehensive and excellent reviews on plant proteomics have dealt with the technologies and their use in various fields in plants [1–7]. In this chapter, first we review the currently available proteomic techniques that are applicable in defense proteome. This is followed by an overview of the results from proteomic studies revealed from plant–pathogen interactions. Finally, perspectives on the future of proteomics related to pathogen responsive proteins and related tools are discussed.

40.2 BRIEF BIBLIOGRAPHIC REVIEW

Recently, a number of proteomic approaches have been applied to understand the plant–pathogen interactions and defense signaling. Compared to the number of publications for plant proteome, the use of proteomic approaches appears rather limited in the field of plant pathogen interaction. A few proteome studies for plant defense response against pathogen have been published (Table 40.1). Among them proteomics has been documented for a differential display of defense responsive proteins, whose expression levels were changed by treatment of fungal elicitors [8, 9] and pathogen attacks such as rice blast fungus [10, 11], *R. solani* [12], rice yellow mottle virus [13], and bacterial pathogen, *Xoo* disease [14], in tissue cultured cell or whole leaf. Similar results have also been reported to other plants such as wheat [15], maize [16], pea [17], and *Arabidopsis* [18]. A proteomics study of *Arabidopsis* proteins induced by *P. syringae* pv. *tomato* DC3000 (disease-inducing), hrpA (activating basal defenses), DC3000 (avrRpm1), which triggers R-gene-mediated resistance, or mock inoculated (10 mM MgCl₂) has recently been reported at the subcellular levels [19]. These studies have mainly employed a classic 2-DGE coupled with MS analysis using MALDI-TOF or LC-MS/MS to identify proteins.

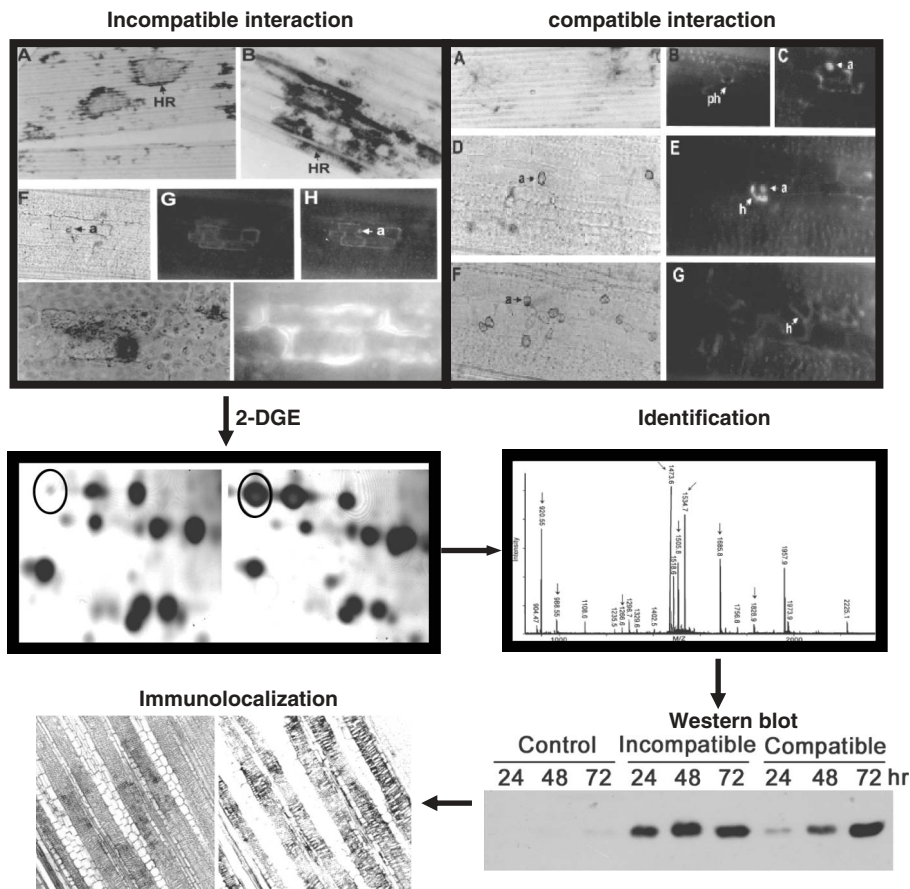


FIGURE 40.1. Strategies for defense proteome analysis. In this approach, protein extracts are prepared from two different samples (e.g., incompatible and compatible interactions) and proteins are resolved by 2-DGE. Proteins that differentially induce by some variable are then selected for identification by MS. Identified proteins remain to be confirmed using Western blot and immunolocalization with antibodies raised against corresponding proteins differentially expressed.

40.3 METHODOLOGY AND STRATEGIES

All proteomic technologies depend on the ability to separate a complex mixture so that 2-DGE is the most widely used tool of protein separation technologies (see also Chapter 2) for plant defense proteome analysis. It is likely that there are a number of problems associated when 2-DGE is used for protein separation and expression profiling technique. For instance, low-abundance proteins present or induced in leaf samples remain difficult to analyze, despite the sensitivity of current instrumentation and high loading capacity of IPGs. Therefore, alternative strategies are required to

TABLE 40.1. Summary of Proteomic Studies in the Field of Plant–Microbe Interactions

Plant	Tissues	Pathogen	Detection	No of Regulated proteins	Identification method	High induced proteins	Reference
Rice	Leaf	Blast fungus	CBB	12	N	Rubisco, PR protein	Konishi et al. [10]
Rice	Suspension cultured cell	Blast fungus (<i>Magnaphothe grisea</i>)	Silver	14	N	PBZ1, OsPR-10, Salt	Kim et al. [8]
Rice	Leaf	Blast fungus	Silver	8	PMF	PBZ1, LTP, RLK	Kim et al. [11]
Rice	Suspension cultured cell	Rice yellow mottle virus	Silver	64	PMF, LC-MS/MS	Salt-induced protein, HSP, SOD	Ventelon-Debout et al. [13]
Rice	Lesion mimic mutant, sdr2	—	CBB	37	LC-MS/MS	PBZ1 and metabolic-related proteins	Tsunezuka et al. [29]
Rice	Leaf	<i>Rhizoctonia solani</i>	SRPGS	21	ESI-Q-TOF	β -glucanase, Putative 3-beta hydroxyl steroid dehydrogenase/ isomerase	Lee et al. [12]
Rice	Leaf	<i>Xanthomonas oryzae pv. oryzae</i>	CBB	20	N	PBZ1 and Rubisco fragmentation, PR-5	Mahmood et al. [14]

Rice	Lesion mimic mutant, blm	—	CBB	26	N and LC-MS/MS	OsPR-5 and metabolic-related proteins	Jung et al. [30]
Pea	Leaf	Powdery mildew (<i>Erysiphe pisi</i>)	CBB	77	Combined with PMF	PR 1 and 5 and metabolic proteins	Curto et al. [17]
Wheat	Leaf	Leaf rust (<i>Puccinia tritica</i>)	CBB	32	QqTOF MS/MS	PR and fungal proteins	Rampitsch et al. [15]
Maize	Suspension cultured cell	Fusarium elicitor	CBB	11	PMF, MS/MS	Peroxidase, GAPDH, HSP	Chivasa et al. [16]
Arabidopsis	Suspension cultured cell	Fusarium elicitor	CBB	8	PMF, MS/MS	GSTs	Ndimba et al. [33]
Arabidopsis	Leaf	<i>Pseudomonas syringae</i>	Silver	7	MS/MS	GSTs	Jones et al. [18]
Arabidopsis	Leaf chloroplast and mitochondria enriched	<i>Pseudomonas syringae</i> DC 3000 hrpA (activating basal defenses and avrRpm1)	Silver, colloidal CBB	52	LC-MS/MS	PRs, GST, metabolic-related proteins	Jones et al. [19]
Total				389			

N, N-terminal sequencing; I, internal amino acid sequencing.

enrich these proteins prior to 2-DGE analysis. In order to increase the detection of certain proteins such as membrane proteins and low-abundance proteins, a number of methods have been developed to fractionate proteins into sub-proteomes based on biochemical, biophysical, and cellular properties. These methods include: (i) PEG fractionation to enrich high-abundance proteins like RuBisCO into a certain fraction from plant leaf [20]; (ii) sequential extraction with a series of reagents based on differential protein solubility [21, 22]; (iii) subcellular proteome for compartments including chloroplasts, mitochondria, nuclei, the extracellular matrix proteome [16]; and the secreted proteome [23]; (iv) LCM for specific tissues [24, 25]; and (v) 2D-DIGE [26, 27]. These procedures reciprocally combined together can greatly improve detection of low-abundance proteins.

PEG Fractionation

For example, plant tissues (such as leaf) that have high content of RuBisCO, the most abundant rice leaf protein, require prefractionation prior to 2-DGE. Differential fractionation of RuBisCO into a certain fraction before loading on the IEF gel will become an important means to expand proteome studies. Kim et al. [20] described the extraction and fractionation technique that utilizes the fractionation power of PEG (MW 4000) in combination with Mg/NP-40 extraction buffer. Rice proteins were extracted with Mg/NP-40 extraction buffer. The Mg/NP-40 buffer extract was further fractionated with PEG into three fractions: 10% PEG and 20% PEG precipitants and the final supernatant fraction that was precipitated with acetone. RuBisCO was enriched in the 20% PEG precipitant. Enrichment was again confirmed by immunoblotting to confirm a validity of this method. This fractionation technique analyzed at least 2600 well-separated protein spots and exhibited less than 1.2% of noticeable overlapping spots (Figure 40.2A). This extraction and fractionation technique is highly effective in rice, and it should be applicable to other plants and will provide more in-depth comparison of plant defense proteome. Figure 40.2B also shows that PEG fractionation combined with alternative methods like basic and giant IEF gel (30 cm) may expand the scope in systemic analyses for plant defense proteome.

Sequential Extraction Methods

Alternative fractionation approaches are based on differential detergent extraction methods that enable simple fractionation of the total proteome into distinct subcellular fractions—for example, cytosolic and membrane proteins. 2-DGE techniques have been aimed at improving and maintaining protein solubility by sequential three differential extraction buffers [28] using chaotropes such as thiourea, detergents such as CHAPS, and reducing agents such as tributyl phosphine (TBP). These methods improved solubility of membrane proteins by the employing reagents with stronger solubilization power instead of traditional solubilizing solutions made of urea, CHAPS, DTT, and carrier ampholytes. Recently a procedure based on the differential solubilization of membrane proteins using C/M mixtures was introduced to extract and concentrate the most hydrophobic proteins, while more hydrophilic proteins were

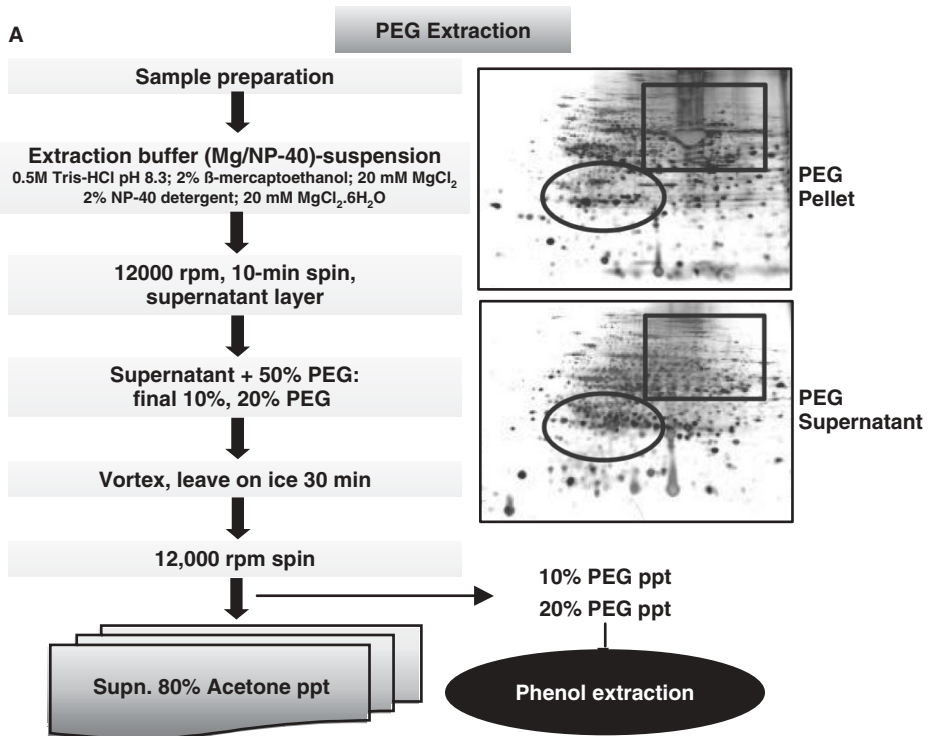


FIGURE 40.2. Overview for experimental application of PEG fractionation studies in plant proteomics. Proteins are extracted from biological samples with Mg/NP extraction buffer and fractionated with PEG. PEG precipitated proteins (PEG ppt) are re-extracted with phenol. PEG supernatant and their pellet fractions show a distinct protein profile (**A**). RuBisCO (rectangles) is fractionated into the pellet. The circles indicate the enrichment of LMW proteins seen in the supernatant. PEG fractionation combined with neutral, basic, and giant IEF gels can be applied into the defense proteome with LMM and leaf inoculated with pathogen including fungi, virus, and bacteria (**B**).

excluded. Santoni et al. [21, 22] evaluated four different extraction procedures (Triton X-100, Triton X-114, carbonate treatment, C/M treatment) for hydrophobic protein isolation and six different LBs [7 M urea, 2 M thiourea, 0.5% (v/v) Triton X-100, 1.2% (v/v) pharmalytes (pH 3–10), 20 mM DTT, and 2% (w/v) of detergent (ØC5, ØC6, ØC7, C8Ø, and ASB14) or 4% (w/v) CHAPS] for solubilization of the isolated hydrophobic proteins. These improved techniques can be useful in applying defense proteomes because the success of the 2-DGE technology for proteome analysis will rely on the ability to display more rare proteins or membrane proteins.

LCM

Another form of sample prefractionation can be achieved at the cellular level. Pathogen-invaded tissues contain mixed populations of cells, with variable proportions

B

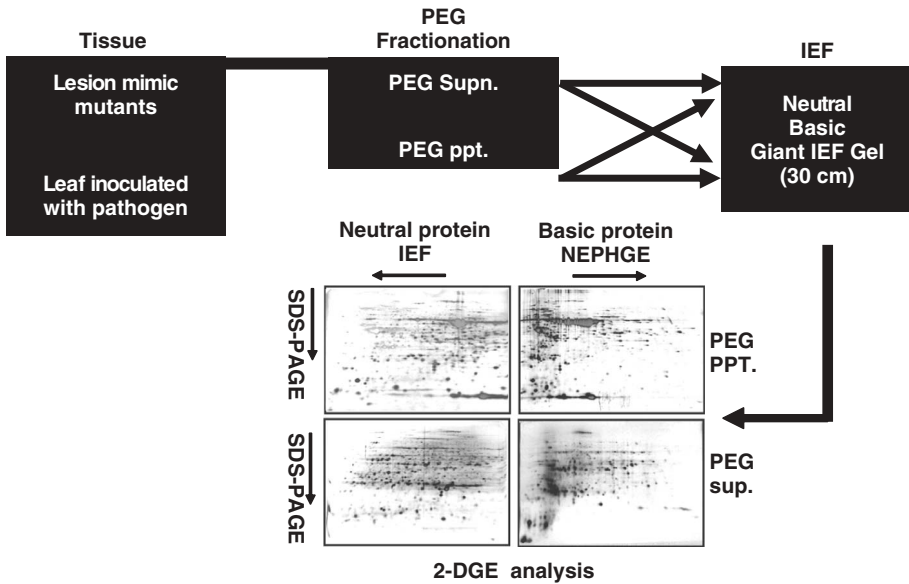


FIGURE 40.2. (Continued)

of diseased and normal cells, resulting in a situation whereby differences of protein expression arising from the disease state could be hindered by the heterogeneity of the sample. LCM allows precise dissection, so that it can be successfully employed to obtain a pure collection of individual cells from a tissue section. Briefly, a slide of stained tissue is placed under a microscope, and a transfer film is placed onto the tissue. The area of interest is centered in the field of vision, and a laser is then fired to melt the film, which expands into the target tissue and solidifies as it rapidly cools. However, LCM is a highly laborious procedure that yields limited amounts of material and thus it is not suitable for screening large numbers of samples. In addition, it remains that the preparation of tissue for microdissection such as the choice of fixative and the staining procedure that may significantly influence both the quantity and quality of the retrieved material will be improved. Nevertheless, it will be greatly used in comparisons of defense proteome; and identified proteins obtained by LCM combined with 2-DGE and nHPLC in conjunction with MS/MS can be confirmed by other analytical techniques, such as immunohistochemistry.

Differential Proteomics using 2D-DIGE

Usually, CBB or silver staining has been utilized to visualize the spots in 2D gel for comparative studies. Proteomics analysis utilizing 2-DGE to determine differential protein expression is frequently criticized due to the time-consuming and laborious processes of image analysis. These processes are complicated by gel-to-gel variations that can affect spot positions, leading to problems with FPs/FNs. Recently, DIGE

using fluorescent dyes to label protein samples prior to 2-DGE, the DIGE has allowed the means to be compared two samples co-separated in the same gel [26, 27]. The dyes used are reported to have enhanced sensitivity in comparison with other commonly used protein stains, and they are believed to be compatible with MS analysis. Thus, DIGE allows very easy comparisons because the samples are run together, and it eliminates the problem of gel-to-gel variations. The protein extracts, for example, one control and one treated, are labeled with different fluorescent dyes (e.g. Cy2, Cy3), then combined and separated by the same 2-DGE. After separation by 2-DGE, scanning with different lasers detects the proteins. The images are then merged, and differences between them can be determined using image analysis software. Even though this approach, however, is still limited by all the problems associated with 2D gels, the DIGE method will be applied successfully to profile biochemical changes within plants treated with a variety of environmental stimuli like pathogen and hormones.

Protein Identification and Data Analysis

Protein identification is now an essential part of almost every proteomics experiment. Some of the studies of plant defense proteome incorporated N-terminal sequencing or sequencing of proteolytic fragments from spots by Edman degradation to generate sequences for database searches [8, 10]. Currently, however, the most commonly used identification method is MS (see also Chapter 3). Software is available to use the two different types of data generated by mass spectrometers to search sequence databases. Using the programs Mascot (www.matrixsciences.com), ProFound (<http://prowl.rockefeller.edu/>), and PeptIdent (www.expasy.org/tools/peptident.html), the peptide mass profiles produced by MS can be analyzed. However, improved software that allows for reliable automated identification of proteins from mass spectrometric data will be required because the analysis of data generated in a proteomics study remains the major cause of errors for many research groups.

40.4 EXPERIMENTAL RESULTS AND APPLICATIONS

Differential Proteome Profiles during PMIs

Rice

Fungi. In rice, it is likely that there are many reports for defense proteome in response to bacteria and fungus compared to the other crop plants due to the complete sequence information of rice genome. Rice blast fungus, *M. grisea*, is a causal agent of rice blast disease, one of the most serious pathogens in rice plant (*O. sativa* L.) in most rice-growing regions of the world. By the use of 2-DGE, protein profiles of the rice leaf and suspension cell after inoculation of *M. grisea* were analyzed during a time course of experiment. Kim et al. [8, 11] performed the first systematic analysis on rice and rice blast fungus interactions using proteomic approach. Kim et al. [8] initially reported the study on the proteomic analysis of rice suspension-cultured cells

inoculated with rice blast fungus or treated with elicitors to identify pathogen-induced proteins. A total of 14 spots were induced or increased. Out of the 14 spots, 12 proteins were identified by either N-terminal or internal sequencing: OsPR10, isoflavone reductase-like protein, β -glucosidase and putative OsRLK, six isoforms of PBZ1 (an OsPR10 family member), and 2 isoforms of SalT. They also applied their study to comprehensive analysis of *M. grisea*-responsive proteins from rice (cv. Jinheung) leaves by use of the PEG (15% w/v) fractionation technique, previously developed to remove RuBisCO into the pellet fraction [11]. A number of spots differentially expressed in response to the infection were identified by MALDI-TOF. The identified proteins were two RLKs, glucanase 1, glucanase 2, and POX 22.3, PBZ1, and OsPR10. Results confirmed using Northern/Western blot analyses revealed that their induction of TLP, OsRLK, PBZ1, and OsPR10 was faster and higher in the incompatible interactions than the compatible ones, implying that early and high induction of these genes may provide host plants to defend themselves from pathogen attack. Two rice PR10 family members, PBZ1 and OsPR10, were subjected to localization studies *in planta*. Interestingly, PBZ1 was localized in the mesophyll cells under the attachment site of the appressorium, which is an indispensable special unicellular fungal structure to successfully infect the rice plant, whereas OsPR10 was present in the vascular tissues, implying that the temporal, spatial, and quantitative differences in expression may provide the host cells with a leading edge in its defense against pathogens.

Sheath blight disease caused by *R. solani* can serve in future studies of rice–fungal interactions because little is known about the host response to the pathogen infection at the protein levels. Lee et al. [12] reported the first proteomic analysis combined with a genetic analysis on the rice response to the fungus. 2-DGE followed by SYPRO Ruby staining and MS/MS analysis using ESI-QTOF-MS was subjected to detect and identify proteins differentially expressed. Eleven of 14 identified proteins (putative 3- β hydroxysteroid dehydrogenase/isomerase protein, stromal APX, putative alpha 1 subunit of 20 S proteasome, RuBisCO activase, RuBisCO large chain, putative chaperonin 60, β precursor 14–3–3-like protein, endosperm luminal Bip, two putative chitinase, and GAPDH A subunit) were induced only in the resistant line against *R. solani*. They can be categorized into the presumed functional groups such as antifungal activity, signal transduction, energy metabolism, photosynthesis, molecular chaperone, proteolysis, and antioxidants. It was proposed that these proteins might be involved in simultaneous induction of proteins through multiple defense pathways following *R. solani* infection. Proteomics combined with genetic analysis will be able to provide a clue for explaining QTLs.

Bacteria. A proteomics approach was performed to identify proteins from the rice leaf blade caused by *Xoo* that were differentially expressed in response to the inoculation [14]. From cytosolic and membrane fraction, 336 proteins were analyzed and 20 were differentially expressed in response to bacterial inoculation. 2-DGE analysis clearly demonstrated that four defense-related proteins such as OsPR5, PBZ1, SOD, and peroxiredoxin were commonly up-regulated by the inoculation of incompatible (T7174) and compatible (Xo7435) race. The induction of OsPR5 and PBZ1

was more rapid and higher in incompatible interactions than the compatible ones. This is in agreement with the result of the rice blast fungus responsive proteome [11]. Therefore, it is likely that the induction of OsPR5 and PBZ1 is a common response to pathogen invasion from fungus to bacteria, implying that it will be helpful in studying the resistance as a marker.

Virus. Although a large number of studies have been published on viral diseases to date, little is known about information of the proteome during host–virus interactions. Recently, protein deregulations and their assignments to early/late and host compatibility in rice yellow mottle virus (RYMV)–rice interactions were studied using a proteomics approach [13]. The RYMV, which is very destructive for rice production and is a member of the *Tobamovirus* genera, is a monopartite positive RNA strand virus containing about 4450 nucleotides and is quite simple, because it is composed of four ORFs. RYMV is considered as a model in rice to study the molecular mechanisms of rice–virus interactions. Analysis of these differentially regulated spots by MALDI–TOF–MS and LC–QTOF–MS/MS resulted in the detection of 32 hypothetical new proteins: 19 for IR64 (Indica; highly susceptible to RYMV) and 13 for Azucena (Japonica; partly resistant to RYMV). HSP70, dehydrin, RAB25, salt-stress-induced protein, PR10a, ethylene-inducible protein, and chaperonin 10 kDa, which are related to stress, metabolism, and translocation, were induced in abundance in IR64, but not found in Azucena, by RYMV, suggesting that RYMV infection also triggers common components of defense/stress pathways like fungus and bacterial interaction. This result can provide further insight into the virus-induced responses and the basis of resistance and/or susceptibility in rice plants.

Pea Curto et al. [17] analyzed changes in the leaf proteome of two pea genotypes differing in their resistance phenotype against powdery mildew disease, *E. pisi*, by a combination of 2-DGE and MALDI–TOF/TOF–MS. Leaf proteins from control, noninoculated and inoculated susceptible (Messire) and resistant (J12480) plants were resolved by 2-DGE, with IEF in the pH 5–8 range. Seventy-seven proteins were identified. The identified proteins mainly belong to three functional categories: photosynthesis, carbohydrate catabolism, and stress and defense responses. From data, a number of the proteins identified correspond to typical PR proteins (PR1, PR5, etc.) that accumulated to high levels in response to pathogen challenge. The antioxidant proteins (such as POX and SOD) that are involved in tolerance to oxidative stress caused by ROS were identified. This is the first report providing an overview of the *P. sativum*–*E. pisi* interaction and shedding light on the putative mechanisms involved in resistance.

Wheat Wheat leaf rust is caused by the basidiomycete fungus *P. triticina* and is a common disease in wheat. The first study on both wheat infected with a virulent race of leaf rust and the fungus itself was reported to investigate (a) the changes of the proteomes of both host and pathogen during disease development (nine days after inoculation), (b) a susceptible line using 2-DGE (with IEF pH 4–8), and

(c) MS [15]. Wheat cultivar ‘Thatcher’ used is susceptible to nearly all races of *P. triticina*. A total of 32 consistently up-regulated protein spots were analyzed by MALDI–QqTOF–MS/MS, followed by cross-species protein identification due to lack of wheat genome sequence. The unmatched spectra were then sequenced *de novo* and queried against the NCBI nr database using the BLAST tools. Seven are host proteins, 22 are fungal proteins of known or hypothetical function, and three are unknown proteins of putative fungal origin.

LMMs LMMs induce a series of defense responses, including expression of defense-related genes and high accumulation of phytoalexins, indicating that the LMMs are useful materials to study defense signaling and PCD in plants. Therefore, LMMs have been isolated and characterized in many plant species, including *Arabidopsis*, maize, barley, and rice. Results of a proteome analysis of the *cdr2* mutant to study defense signaling in rice were recently reported [29]. In this study, total of 37 proteins were differentially expressed between the wild-type and *cdr2* and identified by MS analysis. *Cdr 2* mutant in rice was previously identified to induce activation of defense-related gene expression and high accumulation of phytoalexins and conferred resistance to rice blast fungus. Of 37 proteins identified, defense-related protein and metabolic enzymes were altered during lesion formation, suggesting that the cell death mechanism occurring in the *cdr2* mutant is greatly similar to those of pathogen attack. More recently, a unique rice blast lesion mimic (*blm*) mutant was utilized to analyze differential expression of proteins whose expression levels were altered using immunoblotting and 2-DGE analysis [30]. Many PR proteins (OsPR5 and 10), phytoalexin biosynthesis-related protein (NOMT), and three oxidative-stress-related proteins (catalase, APX, and SOD) were up-regulated in the *blm* mutant. Interestingly, the OsPR5 protein was stained with Pro-Q DPS, suggesting that the OsPR5 is a putative phosphoprotein whose function is unknown. These results support that the comprehensive proteome analysis with LMMs may provide new insights to understand signaling networks involved in defense signaling mechanism in plants.

Application of LCM in Defense Response Proteome As described above, LCM has been applied to analyze plant tissues in a number of studies because it allows analysis of specific cell types from complex tissues. LCM can be applied to study plant–fungi interactions. Howard et al. [31] used LCM to capture the appressorium, a specialized infection structure of rice blast fungus. With the expression of GFP in the cytosol of *M. grisea*, infection hyphae within a single fixed barley epidermal cell were visualized and isolated for subsequent genetic analysis. LCM can also be used to profile proteins from normal and pathogen-inoculated cells (HR cell) *in situ* of the plant (Figure 40.3). After collection of HR cell induced by a pathogen from rice leaf by LCM, for example, proteins were extracted and subjected to protein analysis by the use of the classical 2-DGE combined with LC–MS/MS. Future directions for LCM include analyzing many different samples and pathogen target tissues or cells. Therefore, it is clear that LCM will offer a potential means for deep sampling of a fungal pathogen responsive proteome during the infection process.

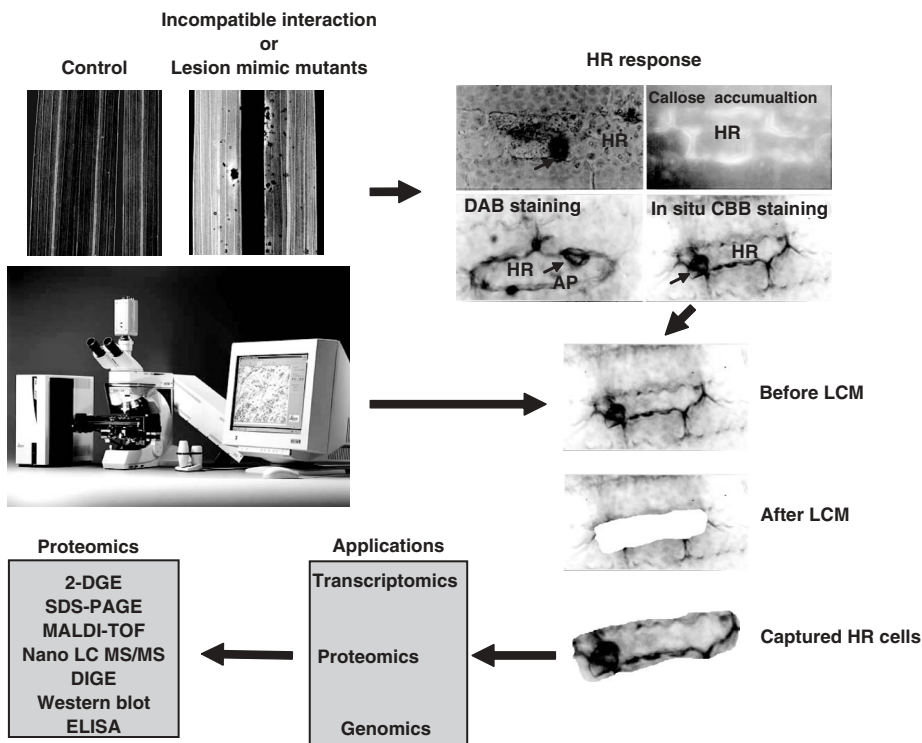


FIGURE 40.3. Overview for application of LCM in plant-microbe interactions.

Subcellular Proteome in Defense and Stress Responses

Chloroplasts and Mitochondria Subcellular fractionation and purification of organelles have provided attractive advantages to low-abundance protein separation in proteomic analysis. The proteomes of some organelles such as chloroplast, mitochondria, nuclei, and other subcellular compartments were studied in plants. Of the organelles, the chloroplast and the mitochondria proteomes are by far the most comprehensively studied. However, these organelles were not investigated in the plant defense proteome. The resistance response against pathogen attack usually triggers a rapid HR that is tightly associated with rapid changes in calcium influx, generation of ROS, expression of PR proteins, production of phytoalexins, and cross-linking of components of the cell wall. Both chloroplasts and mitochondria are additional potential sources of ROS and possess well-developed mechanisms to rapidly detoxify free radicals. Therefore, changes to the proteome of chloroplasts and mitochondria are of particular interest because this organelle was implicated in the defense response in diverse ways.

Recently, it has been reported that subcellular proteome analysis in *A. thaliana* leaves during responses to challenge by *P. syringae* pv. *tomato* DC3000 (disease-inducing), hrpA (activating basal defenses), and DC3000 (avrRpm1, which triggers

R-gene-mediated resistance), or mock inoculated (10 mM MgCl_2) were analyzed using 2-DGE [19]. Total soluble protein-, chloroplast-, and mitochondria-enriched fractions were isolated, and proteins were verified by LC-MS/MS. Seventy-three differential spots, representing 52 unique proteins, were successfully identified. Identified proteins showed significant change after bacterial challenges were grouped as two main functional groups: defense-related antioxidants and metabolic enzymes. However, this approach has not been subjected to other plants like rice. So, as a next step, the study for the subcellular proteomes will be an essential approach to find new low-abundance proteins involved in defense proteome. In particular, this approach will provide a clue regarding how communication between organelles occurs when plant exposed to pathogen and other stresses.

Extracellular Matrix It is known that the extracellular matrix (ECM) where complicated biological events take place plays a variety of roles in growth, development, and defense against environmental stresses. More recently, evidence has been obtained that the ECM has an important function during plant defense, such as reinforcement of the cell wall, antimicrobial activity by PR proteins (chitinases and glucanase, thionins, defensins, LTPs), and signaling cascade. ECM containing secreted proteins has a dynamic role serving as the first mediator of cell-cell interactions in which cells perceive extracellular signals released from extracellular environmental stresses in diverse ways. Before the final response is induced, plants must perceive such stimuli, transfer the signals, and transduce them to the site of reaction. It has been suggested that the ECM is involved in the response as well as in the perception and transduction of environmental signals with the PM. In proteome analysis, the plant extracellular matrix/secretome is more studied than those of other subcellular compartments in plant defense proteome. It is likely that the liquid cell culture medium is useful for proteome analysis of ECM due to easy separation from the cells without cell disruption. For example, cultured BY2 cells were used to examine the extracellular proteins of tobacco (*N. tabacum*) cells. Several PR proteins (e.g., ascorbate oxidase, POX, and *exo*-glucanase) were found by N-terminal amino acid sequencing [32]. As a model plant, *Arabidopsis* is a wonderfully powerful material due to easy combination of genetic, genomic, and proteomics approach. Researchers have reported a proteomic approach in the ECM of *A. thaliana* cell suspension cultures following treatments with a variety of elicitors: two fungal pathogen elicitors, chitosan, and extracts of *F. moniliforme* to investigate proteins changes [33, 34]. Using a combination of 2-DGE, immunoblotting with a pTyr-specific antibody, and MALDI-TOF-MS, researchers identified pathogen elicitor responsive proteins, lectin RLK, a putative endochitinase, a XEG, and a POX in the culture filtrate extracts. From data, lectin RLK, a putative endochitinase, and a XEG were phosphorylated on tyrosine residues within their proteins, implying that extracellular phosphorylation network can be occurred to communicate between cells intercellularly. Oh et al. [23] isolated and identified proteins involved in a variety of cellular functions such as signaling and regulation, catalysis, protein degradation, protein protection, and structure in the culture medium of an *Arabidopsis* suspension cell. They isolated a total of 13 proteins whose levels are changed in the secretome by treatment of SA, which is involved in

the defense process. Of these, GDSL motif lipase was further investigated in disease resistance to the fungus, *A. brassicicola*. Some of these proteins were also found in the ECM, such as an arabinogalactan protein known to be associated with cellular differentiation and signaling cascades. So far, however, proteomic analysis of ECM or secretome in response of hormones, pathogen, and fungal elicitor has not been extensively studied in whole plants. Preliminarily, we investigated the secreted proteome induced by a rice blast fungal pathogen in a rice suspension cultured system, resulting in the fact that 14 proteins were significantly induced in suspension cultured medium. Normal and pathogen-invaded cells were analyzed by 2-DGE (Figure 40.4). This approach allows the concentration of proteins released into the liquid cultured medium. By focusing only on the secreted proteins found in the tissue culture media, 2-DGE followed by MS can enable us to identify the secretomes. Therefore, ECM proteome may be one of the solutions for understanding on the complex defense mechanisms.

40.5 CONCLUSIONS

The rapid improvement of MS, protein separation techniques, and novel methods like sample preparation and comparison has widened the spectra of potential applications for proteomics. In addition, these technologies offer the opportunity to analyze protein expression profiles and to analyze the proteins involved in signaling pathways. At present, the plant defense proteome has been directed at a simple approach. In the future, it can be directed to an entire set of approaches all along the pathogenic signaling pathway. A suite of advanced technologies have recently been describing with coupling of these methods to high-resolution proteome analysis methods, such as nLC-MS/MS, which promises the emergence of qualitative studies of whole proteomes. Different approaches to protein isolation selected for a distinct subset of the proteome (e.g., acid, basic, LMW and HMW), along with different separation and detection methods, have provided unique advantages in plant proteomics. As we described above, prefractionation, sequential extraction, LCM, 2D-DIGE, and MS highly offer a valuable result in plant and can achieve a high efficiency of low-abundance protein. Currently, a number of defense responsive proteins have been identified by proteome studies, and almost all of these are employing similar approaches like classic 2-DGE and protein extraction methods in *Arabidopsis*, rice, pea, wheat, and maize. Not as expected, however, the greatest number of proteins identified so far typically belongs to the PR family proteins, ROS-related proteins, and metabolic enzymes, implying that the demand for enriched protein samples or pathogen target-specific tissues (HR and vascular bundle) using technologies described above will continue to rise. Interestingly, PR proteins and ROS-related proteins are common factors that are induced by all biotic stresses including bacteria and fungus, indicating that these could be used as defense/stress-related marker proteins. It is notable that PBZ1, APX, and SOD induced by pathogens are known to have many isoforms showing different pIs on 2D gels in plants inoculated with pathogen and in LMM [8, 30]. However, it still remains as to how these different isoforms were regulated on their biological activity.

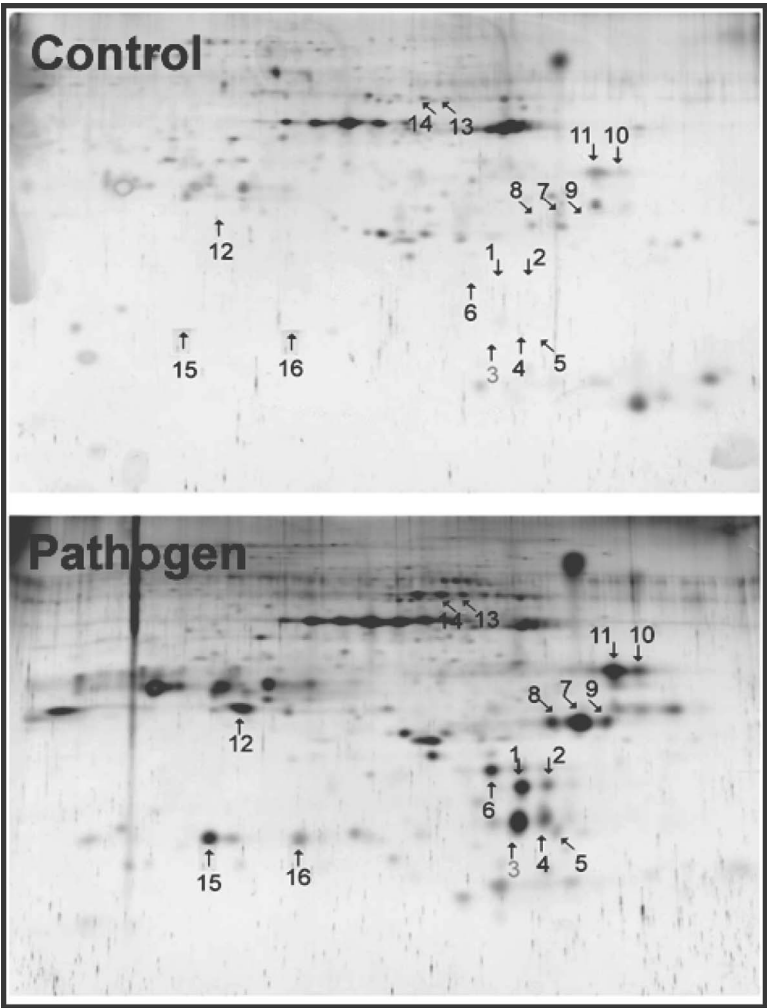


FIGURE 40.4. 2-DGE analyses of secreted proteome from rice suspension cultured cell incubated *in vitro* in the absence (control) or presence (pathogen) of rice blast fungus. Black arrows indicate the differentially induced proteins compared to the noninoculated sample (control). Gels are stained with silver nitrate.

40.6 FIVE-YEAR VIEWPOINT

One of the areas where it has the greatest potential is in the discovery of new defense proteins involved in plant disease resistant. Perhaps, in the future, the most important issue that needs to be considered in the proteomics analysis of pathogen-target-specific tissues is the ultimate goal of sample preparation. LCM enables pathogen-target-specific and homogenous cell types from thin tissue sections to be isolated. Thus, we

predict that tissue-specific global proteome analysis of both host and pathogen, using LCM combined with nLC-MS/MS and 2-DGE, can provide valuable insights into the mechanisms that allow successful infection and survival in host tissues. Large-scale transcriptional analyses with microarrays have been widely utilized, the analysis of which has revealed coordinated changes in hundreds of transcripts altered by environmental stresses. As mentioned above, however, it is clear that proteomics changes cannot be directly predicted from changes in the transcriptome. Furthermore, proteins and transcripts interact with other molecule in specific ways *in planta*, resulting in determination of the regulation, expression, activity, and stability of specific mRNA and protein. Therefore, combination with transcriptomics and proteomics can also provide a useful approach to gain a better understanding of the function of the defense responsive proteins. In summary each combination with classic analytical techniques such as 2-DGE with prefractionation techniques and MS will also represent a considerable progress in cataloguing and quantifying proteins present in plants inoculated with pathogen, LMM tissue, and subcellular compartments in both normal and diseased physiological states. Finally, there are still many alternative approaches we did not mention above for proteomic analysis in plant defense proteome. Immunoprecipitation for identification of PPIs, as well as immunoblotting/immunolocalization for characterization of antibody specificity or identification of PTMs of proteins, can be considered useful for characterization of proteins identified. With a variety of proteomics methods mentioned above all combined together, ultimately the investigation of the plant defense proteome will lead to new insights in the near future.

ACKNOWLEDGMENT

This work was financially supported by the Environmental Biotechnology Research Center (EBRC) and Crop Functional Genomics Center (CFGC).

REFERENCES

1. Agrawal GK, et al. 2005. *J Chromatogr B* **815**:109–123.
2. Agrawal GK, et al. 2005. *J Chromatogr B* **815**:125–136.
3. Agrawal GK, et al. 2005. *J Chromatogr B* **815**:137–145.
4. Agrawal GK, et al. 2006. *Proteomics* **6**:5549–5576.
5. Agrawal GK and Rakwal R. 2006. *Mass Spectrom Rev* **25**:1–53.
6. Chen S and Harmon AC. 2006. *Proteomics* **6**:5504–5516.
7. Rossignol M, et al. 2006. *Proteomics* **6**:5529–5548.
8. Kim ST, et al. 2003. *Proteomics* **3**:2368–2378.
9. Fujiwara M, et al. 2006. *Plant Physiol* **140**:734–745.
10. Konishi H, Ishiguro K, and Komatsu S. 2001. *Proteomics* **1**:1162–1171.
11. Kim ST, et al. 2004. *Proteomics* **4**:3569–3578.
12. Lee J, et al. 2006. *Mol Plant Pathol* **7**:405–416.

13. Ventelon-Debout M, et al. 2004. *Proteomics* **4**:216–225.
14. Mahmood T, et al. 2006. *Proteomics* **6**:6053–6065.
15. Rampitsch C, et al. 2006. *Proteomics* **6**:1897–1907.
16. Chivasa S, et al. 2005. *Proteomics* **5**:4894–4904.
17. Curto M, et al. 2006. *Proteomics* **6**:163–174.
18. Jones AM, et al. 2004. *Phytochemistry* **65**:1805–1816.
19. Jones AM, et al. 2006. *Plant Physiol* **142**:1603–1620.
20. Kim ST, et al. 2001. *Electrophoresis* **22**:2103–2109.
21. Santoni V, et al. 2000. *Electrophoresis* **21**:3329–3344.
22. Santoni V, Molly M and Rabilloud T. 2000. *Electrophoresis* **21**:1054–1070.
23. Oh IS, et al. 2005. *Plant Cell* **17**:2832–2847.
24. Kerk NM, et al. 2003. *Plant Physiol* **132**:27–35.
25. Schad M, et al. 2005. *Electrophoresis* **26**:2729–2738.
26. Tonge R, et al. 2001. *Proteomics* **1**:377–396.
27. Alban A, et al. 2003. *Proteomics* **3**:36–44.
28. Molloy MP, et al. 1998. *Electrophoresis* **19**:837–844.
29. Tsunazuka H, et al. 2005. *Mol Plant Microbe Interact* **18**:52–59.
30. Jung YH, et al. 2006. *J Proteome Res* **5**:2586–2598.
31. Howard RJ, Bourett TM, and Czymmek KJ. 2004. *Microsc Microanal* **10**:218–219.
32. Okushima Y, et al. 2000. *Plant Mol Biol* **42**:479–488.
33. Ndimba BK, et al. 2003. *Proteomics* **3**:1047–1059.
34. Chivasa S, et al. 2006. *J Exp Bot* **57**:1553–1562.

PROTEOME ANALYSIS OF CELLULAR RESPONSES TO ABIOTIC STRESSES IN PLANTS

Hans-Peter Mock and Andrea Matros

41.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

The following contribution will summarize results from a range of publications, making use of different proteome approaches to study cellular responses of plants toward abiotic stress factors. Plants have to deal with unfavorable environmental conditions imposed by a single or several stresses at most stages of development and have to respond adequately to ensure survival (Figure 41.1). In addition, crop plant productivity is severely influenced, and abiotic stresses lead to considerable yield losses. Therefore, emerging proteome approaches were rapidly applied in this field of plant biology, as shown by the large number of publications. It can be expected that proteome studies will significantly contribute to gain further insights into the cellular networks underlying plant stress defense. These results will have a high impact on future plant breeding and biotechnology. We will first give a short overview on previous research on abiotic stresses in plants, and we will then deal with the proteomic approaches published on different stress factors. We will also summarize results obtained by transcript profiling and compare the advantages of both “omics” techniques to obtain a deeper insight into abiotic stress defense mechanisms. Finally, we

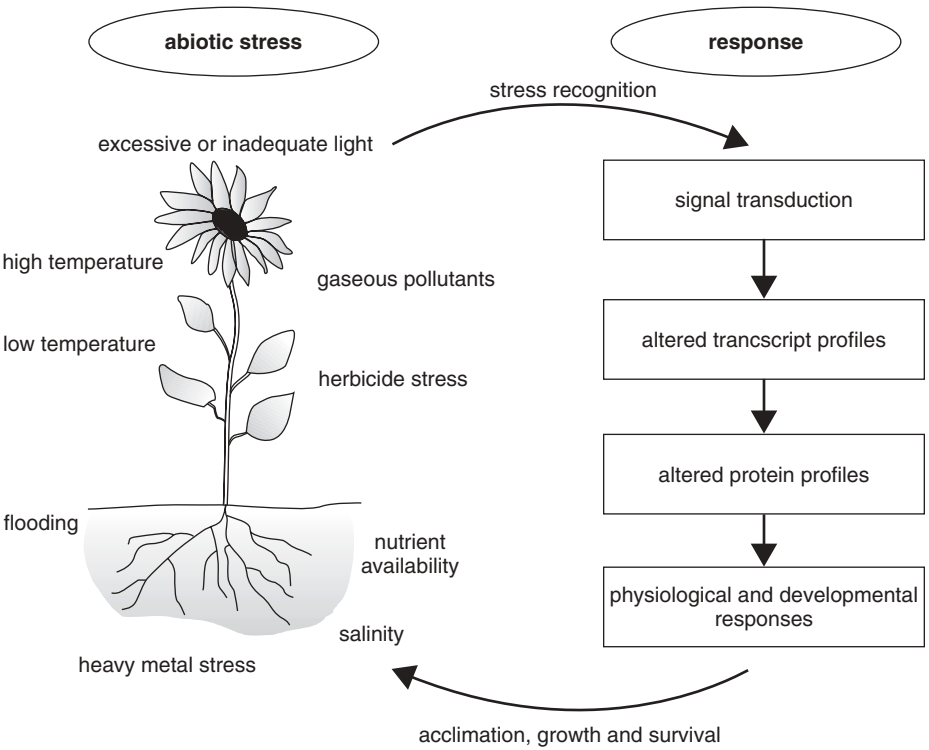


FIGURE 41.1. Abiotic environmental factors can act singly or in concert to strongly impair plant growth and productivity, thereby driving plants to respond adequately to ensure their survival. Plant responses leading to stress tolerance include constitutive resistance mechanisms besides physiological alterations after stress recognition, resulting in acclimation to an unfavorable environment. As a result, stress recognition is followed by transduction of the environmental signal throughout the plant. Reactions on the cellular level are changed gene expression values and alterations on the protein level, which in turn can affect metabolism and development of the whole plant.

will give a perspective for work in this field for the near future, and we will argue that research will greatly benefit from combined approaches also involving defined genetic resources.

41.2 SUMMARY OF PREVIOUS RESEARCH ON ABIOTIC STRESSES IN PLANTS

A number of unfavorable environmental factors can act singly or in concert to strongly impair plant growth and productivity; these include the suboptimal availability of nutrients, excessive or inadequate light, lack of oxygen, lack of water, high or low temperature, wind, and chemical pollution. These abiotic stress factors have a serious

impact on plant crop productivity worldwide. It has been calculated that yield losses can amount up to 80% of the record yield due to abiotic stresses in major crop plants, depending on the species [1]. Therefore, basic and applied research has been directed toward improved knowledge on stress defense responses and generation of plants with higher stress tolerance. Geneticists have screened for accessions of major crop species with improved stress tolerance [see reviews in reference 2]. Oxidative stress is a common feature associated with many stress factors. Biochemical and molecular analysis has elucidated enzymatic and nonenzymatic components coping with excess reactive oxygen species, such as SOD, APX, catalase and GR and LMW antioxidants such as ascorbate, tocopherol, carotenoids, or flavonoids [see reviews in reference 3].

Because drought and salinity are the most commonplace environmental limitations to plant productivity, a number of studies have addressed these factors, and we will start our chapter with a summary of related publications. Although under field conditions plants frequently have to cope with a combination of stresses, we will structure the review by summarizing work by single factors and then address commonalities in the responses toward different factors. We will also briefly summarize studies on transcript profiling of abiotic stress defense responses and compare the data obtained with the proteomics approaches.

Proteome Analysis of Cellular Responses to Different Abiotic Stress Factors

Plant growth and development are strongly regulated by a combination of biotic and abiotic environmental factors, which, due to the sessile way of living abnormalities from the physiological optimum of one or a number of environmental stimuli, can cause dramatic effects on plant growth, development, and productivity. To cope with these stresses, plants have evolved a number of defensive strategies linking perception, transduction, and response reactions in a complex network. Single factors of those reactions have been investigated for a couple of years, but our knowledge on the highly sophisticated and interlinked network of responses to single stress factors or multiple stresses is still rather rudimentary. New possibilities to gain insights into that scientific field are studies at the level of the proteome, because it reflects the physiological active part of an organism on the molecular level. Thus, in a large number of reports, proteome approaches were employed to investigate differences in cellular protein expression in response to various abiotic stress factors (Table 41.1).

Drought. Protein changes occurring in maize leaves after dehydration were described in reference 4 by use of two unrelated lines of maize and their hybrid. Changes induced in leaf proteins were studied by 2-DGE and revealed 78 proteins showing significant quantitative differences. Among them, proteins already known to be involved in water stress have been identified (e.g., RAB17, ABA-stress-ripening (ASR) protein, and ferritin) in addition to a number of enzymes involved in basic metabolic pathways such as glycolysis and the Krebs cycle. In addition, increased amounts were observed for a number of proteins involved in phenylpropanoid metabolism, such as COMT, β -glucosidase, cysteine synthase, and glutamate semialdehyde aminotransferase. Thus, activation of the lignin-biosynthesis pathway by

TABLE 41.1. Reports on Proteome Analyses of Cellular Responses to Different Abiotic Stress Factors

Stress Condition	Plant Species	Investigated Material	References
Drought	<i>Zea mays</i> L.	Leaves	4, 5, 9
	<i>Lupinus albus</i>	Stem	6
	<i>Beta vulgaris</i> L.	Leaf	7
	<i>Oryza sativa</i> L.	Leaves, leaf sheaths, seedlings	8, 10, 11
	<i>Quercus ilex</i> L.	Leaf	12
Salinity	<i>Oryza sativa</i> L.	Seedlings, roots, leaves of seedlings, leaf lamina	13–16, 25
	<i>Zea mays</i> L.	Roots and shoots	17
	<i>Suaeda aegyptiaca</i>	Leaf	18
	<i>Pisum sativum</i>	Root	19
	<i>Arabidopsis thaliana</i>	Root microsomal fractions	20, 24
	<i>Nicotiana tabacum</i>	Leaf apoplast	21
	<i>Synechocystis</i> sp.	Periplasmic fractions, plasma membranes	22, 23
Flooding	<i>Oryza sativa</i> L.	Shoots of rice seedlings	83
Nutrient availability	<i>Arabidopsis thaliana</i>	Seedlings, cell suspension	28, 29
	<i>Triticum aestivum</i>	Leaf	30
	<i>Nicotiana tabacum</i>	Cell suspension	31
	<i>Lupinus albus</i>	Leaf apoplast	32
	<i>Picea abies</i> L. Karst.	Suspension-cultured spruce cells	33
Heavy metal stress	<i>Arabidopsis thaliana</i>	Seedlings, cell suspension	34, 35
	<i>Vigna unguiculata</i>	Leaf apoplast	36
	<i>Alyssum lesbiacum</i>	Roots and shoots	37
	<i>Zea mays</i> L.	Seedlings, plantlets	38, 39
Gaseous pollutants	<i>Oryza sativa</i> L.	Leaf	reviewed in reference 40
	<i>Zea mays</i> L.	Leaf	reviewed in reference 40
	<i>Arabidopsis thaliana</i>	Leaf	41
Herbicide stress	<i>Triticum tauschii</i>	Coleoptiles	42
	<i>Vitis vinifera</i> L.	Leaves, roots and shoots of plantlets	43
	<i>Glycine max</i> L.	Leaf	44
Excessive or inadequate light	<i>Nostoc commune</i>	Membrane, cytosolic and extracellular fractions	45
	<i>Zea mays</i> L.	Leaves, roots	46, 47

TABLE 41.1. (Continued)

Stress Condition	Plant Species	Investigated Material	Reference
High temperature	<i>Arabidopsis thaliana</i>	Isolated chloroplasts, luminal fractions, plastoglobules from chloroplasts	48–50
	<i>Capsicum annuum</i>	Plastoglobules from isolated chromoplasts	50
	<i>Triticum aestivum</i>	Grains	51
	<i>Arabidopsis thaliana</i>	Leaf and leaf peroxisomes	52
	<i>Populus euphratica</i>	Leaf	53
Low temperature	<i>Oryza sativa</i> L.	Anthers, seedlings, leaves of seedlings,	54–57
	<i>Arabidopsis thaliana</i>	Plasma membranes, nuclear fractions, leaves	58–60
	<i>Populus tremula</i> L. \times <i>Populus tremuloides</i> <i>Michaux</i>	Leaves	61, 62

water deficit was suggested. Later on, 2-DE profiling together with an analysis of secondary metabolism was applied to elucidate the impact of water deficiency on the lignification of maize leaves [5]. The phenotypic response to water deficit was a reduction in the level of lignin and a shift of the lignification zone toward the leaf base. The expression profiles with respect to candidate proteins were consistent with the morphological and biochemical alterations.

Previously, 2-DGE combined to MS was used to investigate the effect of water deficit on the protein composition of *L. albus* plants [6]. In the early stage of drought stress, significantly increased levels of serine proteases were observed in the lupin stem. At severe water deficit, the accumulation of immunophilin, serine protease, and cysteine protease was reported, which are well-known components of sensing pathways in animals. Furthermore, the simultaneous expression of proteases and protease inhibitors reacting differently to the stress level and to rewatering was found.

The response to drought stress of field-grown sugar beet has been described in reference 7. Analysis of the 2-DGE profiles of leaf extracts from two different cultivars identified 79 gene products whose intensity changed significantly as a consequence of water deficit; of these, 44 were down-regulated, 27 were up-regulated, and eight were not detected in the nonstressed controls. Expression of some of these proteins was also genetically variable. Selected spots were purified for protein identification, revealing proteins that have previously been shown to play a role in the stress response, including the smHSP, cytosolic Cu-ZnSOD, 2-cysteine POX, cyclophilin, NDPK, a nascent polypeptide-associated complex α -chain, and the LSU of RuBisCO.

Prior proteomics analyses based on a number of species have identified a significant degree of inter- and intraspecific genetic variation in the expression of these stress response gene products. Some of these proteins clearly give a physiological advantage under stress conditions, and thus they are simultaneously potential targets for marker-assisted selection and rational candidate genes for the identification of quantitative trait loci.

Cultivar-specific differences in protein expression pattern during drought stress have also been investigated using *O. sativa* L. cv. CT9993 and cv. IR62266 [8]. Forty-two proteins were found to be significantly up- or down-regulated in abundance. Fifteen proteins behaved similarly in the two cultivars, whereas 27 displayed increased or decreased abundance in one line and failed to respond in the other. For only one protein, a chloroplastic Cu–ZnSOD, different directions in regulation of expression during drought stress were observed within the two cultivars. In a similar study, protein responses to water deficit have been investigated in leaves of two inbred lines of maize [9]. Decreased levels were observed for eight proteins, and increased abundance was found for 46 proteins during water deficit. Among the proteins with increased levels of OEE protein 1, three isoforms of β -glucosidase, ABA45, a MDH, a GSAAT, a protein similar to the rice OSR40 protein, a COMT, a cysteine synthase, an ASR-induced protein, an enolase, and a cystatin were identified. For some of them, specific genotypic variations in the shape or the slope of their curve of responses as a function of time were detected (e.g., β -glucosidase, ABA45, OSR40, ASR1, cystatin). Interestingly, some of these differences (e.g., ABA45, OSR40) were shown to be linked to variations in ABA accumulation in the leaves of the two genotypes. Consistent with these findings, drastic changes in the protein profile of rice seedlings by ABA treatment including altered accumulation of RuBisCO and certain defense/stress-related proteins was reported [10]. Furthermore, data revealed remarkable changes in the phosphorylation status of some proteins, along with differential effects on MBP and CDPK activities during ABA treatment. More recently, the protein patterns of leaf sheaths of *O. sativa* cv. Nipponbare and the drought-tolerant cultivar Zhonghua 8 were compared [11]. Changed protein levels were determined for an actin polymerization factor, PS II OEC protein, OEE protein 2, and LHC chain II in the drought-tolerant cultivar Zhonghua 8 relative to the cultivar Nipponbare. The responses to different stresses were also investigated. Actin depolymerizing factor, light harvesting complex chain II, SOD, and a mannose-binding lectin (SalT) were shown to be induced in response to drought and osmotic stress, but which remained unchanged during salt stress, cold stress, and/or ABA treatment, suggesting its involvement in conferring resistance to drought and osmotic stress; however, this remains to be confirmed in new studies

More recently, the effects of drought on the leaf proteome of Holm oak (a forest tree species) have been described. The major physiological changes were the mobilization of storage proteins and carbohydrates, as well as the inhibition of photosynthesis. Other differences at the level of the leaf proteome related to the developmental stage of the leaf and provenance of the tree [12].

Salinity. Indications for dose- and time-dependent responses to salt stress in a tissue-specific manner were given by the investigation of rice seedlings exposed to 50,

100, and 150 mM NaCl for 6–48 h [13]. Cultivation of rice seedlings at low salinity resulted in enhanced accumulation of eight proteins with a maximum expression level at 24 h. In contrast, longer exposure and exposure to NaCl levels higher than 50 mM caused down-regulation of proteins which were initially responsive. Differences in expression levels of affected proteins have also been observed for leaf sheath, leaf blade, and roots. As a result, it was suggested that the induced expression levels of SOD, OEE protein 2, PS II OEC protein, and aldolases in leaf sheath and leaf blade are part of responses of rice against low doses and short-term NaCl stress. Later on, investigation of the rice cultivar Nipponbare revealed an additional number of proteins involved in salt stress responses of rice roots; these proteins included enolase, UDP-glucose pyrophosphorylase, cytochrome c oxidase subunit 6b-1, glutamine synthetase root isozyme, putative nascent polypeptide-associated complex alpha chain, putative splicing factor-like protein, and putative actin binding protein [14]. Salt stress responses on the leaf proteome were also investigated in the third leaf of rice seedlings by using a hydroponic culture system [15]. In this study, 55 differentially expressed spots were detected, with 47 spots showing increased levels when seedlings were grown under 130 mM NaCl. Identified proteins were mainly involved in photosynthetic CO₂ assimilation and photorespiration. Remarkable changes were also observed for markers of oxidative damage, such as induction of APX and lipid peroxidation, and suppression of catalase, which was confirmed in 2-DGE immunoblot and enzymatic activity analysis. More recently, the proteome analysis of short- and long-term responses to salt stress was performed in rice leaf lamina [16]. Exposure to 50 mM NaCl resulted in increased expression of RuBisCO activase and ferritin at 24 h and continued to increase during the following 6 days. One protein, a putative PGK, showed enhanced accumulation with a maximum expression level at 24 h, while others had increased (e.g., SOD) or decreased expression levels (e.g., SAM synthetase) only after 7 days of salt treatment.

Although maize is considered to be strongly salt-sensitive, proteome studies to investigate responses to salt stress in this species are rare. By using an Na⁺-excluding maize inbred line, protein patterns in response to salt stress were analyzed for roots and shoots [17]. Results from this experiment revealed three groups of proteins affected mainly by low salt stress: (i) enzymes involved in protein biosynthesis and protein modifications by kinases, (ii) proteins associated with carbon metabolism, and (iii) enzymes of N metabolism. Because dramatic changes were detected on the protein level without biomass and ion effects even under low salt concentrations, a general avoidance mechanism to prevent damage was suggested for maize rather than clear physiological roles of affected gene products.

Based on 2-DGE separation, the effects of salinity levels on the leaf proteome of the salt-tolerant plant *S. aegyptiaca* have been investigated [18]. Ten-day-old plants were exposed to 0, 150, 300, 450, and 600 mM NaCl and analyzed after 30 days of treatment. Image analysis revealed 102 spots significantly responding to salt treatment when compared to the control. For grouping hierarchical and nonhierarchical statistical methods were used, leading to 12 different expression groups. Of these, 27 proteins were identified including proteins involved in oxidative stress tolerance,

glycine betaine synthesis, cytoskeleton remodeling, photosynthesis, ATP production, protein degradation, cyanide detoxification, and chaperone activities.

A study covering the changes in protein levels of roots from *P. sativum* in response to salinity was presented by reference 19. Major changes in abundance were found for PR10 proteins and antioxidant enzymes such as SOD and NDPK, highlighting the potentially crucial role of these polypeptides in mediating abiotic stress responses. A comprehensive study of the signaling events during salt stress in *A. thaliana* analyzed salt-induced changes in the root microsomal proteome, using 2-DGE combined with protein gel blotting [20]. Ca^{2+} -dependent membrane-binding proteins, designated annexins, were identified as the major signaling components, a result validated by the observation that T-DNA insertion mutants of two isoforms of annexin displayed hypersensitivity to osmotic stress and ABA treatment during germination and early seedling growth.

Changes induced in the response to salinity stress have also been described at the level of the soluble fraction of tobacco leaf apoplastic polypeptides. The expression of two chitinases and a germin-like protein increased significantly, and two LTPs were expressed *de novo* [21]. In a parallel study, salt-stress-induced changes in the periplasmic fractions of *Synechocystis* were analyzed [22] and, more recently, a proteomics approach based on isolated PM was developed [23]. The largest defined group of proteins which consistently shows enhanced expression under salt stress contains the substrate-binding domains of ABC transporters. Regrettably, a substantial number of proteins affected by salinity remain annotated as hypothetical proteins, due to a lack of any sequence similarity to any proteins of known function. The effects of salinity and hyperosmotic stress on *A. thaliana* cell suspension cultures have been investigated using DIGE technology [24]. Products displaying significant changes in abundance included H^+ transporting ATPases, signal transduction-related proteins, transcription/translation-related proteins, detoxifying enzymes, amino acid and purine biosynthesis-related proteins, proteolytic enzymes, smHSPs, carbohydrate metabolism-associated proteins, and proteins of unknown function. In addition, the transient suppression of *de novo* protein synthesis was demonstrated. Consistent with induced levels of H^+ transporting ATPases during salt stress in that study, more recently a mitochondrial ATP synthase 6-kDa subunit gene was identified from a cDNA library prepared from rice roots grown in the presence of NaHCO_3 stress [25]. Analysis of the corresponding mRNA levels in transformed yeast and tobacco protoplasts revealed induction of this gene during stress by sodium carbonate and other sodium salts, and transgenic tobacco overexpressing the ATP synthase 6-kDa subunit gene displayed greater tolerance to salt stress in the seedling stage.

Flooding. Low oxygen availability during flooding stress results in enormous yield losses in numerous crops worldwide; therefore many efforts were made in evaluation of molecular, biochemical, physiological, and genetic factors involved in adaptation to flooding stress [reviewed in references 26 and 27]. Decreased content of total nitrogen in plants is one of the main responses observed during flooding stress and is also reflected by increased activities of key enzymes involved in nitrate reduction and NH_3 assimilation, such as nitrate reductase and alanine aminotransferase. In addition, glycolysis followed by alcoholic fermentation was suggested as

the main ATP-generating pathway during oxygen deficiency, indicated by induction of enzymes of carbohydrate metabolism, glycolysis, and ATP transport in many plant systems under anoxia. One example was reported for shoots of rice seedlings [83]. Thus, proteome analysis of flood-tolerant and flood-sensitive cultivars strongly indicated involvement of sucrose synthase, GAPDH, UDP-glucose-6-dehydrogenase, and asparagine synthetase in flooding tolerance of rice. Thus carbohydrate is considered as the main player in controlling responses to low-oxygen stress.

Nutrient Availability. Mineral nutrient deficiency is one of the most severe abiotic factors threatening agriculture worldwide. In a comparative proteome analysis, proteins differentially expressed by insufficient K^+ nutrition have been analyzed in *A. thaliana* seedlings [28] after treatment for 3 h or 7 days. A large number of proteins with changed expression patterns were determined, among them being putative TFs, protein kinases, and phosphatases, in addition to proteins playing a role in phytohormone biosynthesis/signaling, carbon and energy metabolism as well as proteins involved in regulation of diverse signal transduction pathways, such as 14–3–3 proteins and small G-proteins. Later on, two reference 2D gel maps of major soluble protein fractions from *A. thaliana* cultured in two different media (Gamborg B5 and Murashige and Skoog) were reported in order to determine the effect of different nutrient supply on the soluble proteome [29]. Strongly increased amounts were observed for a number of housekeeping enzymes and stress/defense proteins besides modified expression patterns for enzymes involved in nitrogen and sulfur metabolism, indicating a significant influence of the respective culture medium on the expression pattern of major soluble plant proteins.

Bahrman et al. [30] described a proteome study on the leaf protein extracts of two wheat varieties (Arche and Récital) grown under four nitrogen supply conditions, with the aim to investigate differences in nitrogen utilization. A significant variety effect was detected for 55 protein spots, and a significant nitrogen-treatment effect was observed for 76 protein spots. Among the proteins identified were mainly enzymes involved in carbon fixation and energy production, such as 2,3-biphosphoglycerate-independent PGMA, enolase, phosphoglycerate kinase, FBPase, MDH, and ATP synthase beta subunit. However, because a large number of proteins with differential expression patterns during the varying nitrogen treatments and/or the diverse varieties still remain to be identified, a full picture of responses to differences in nitrogen utilization could not yet be made.

In a more functional approach, the role of alternative oxidase (AOX) in modulating carbon use efficiency and growth during macronutrient (P or N) stress was investigated in tobacco suspension cells [31]. Strong accumulation of AOX was observed in wild-type tobacco cells grown under macronutrient deficiency, an indication for activation of the non-energy-conserving branch of the respiratory electron transport chain. Northern analyses and comparative analysis of the mitochondrial proteome of wild-type cells and transgenic cells lacking AOX accumulation revealed mainly increased levels for proteins strongly related to oxidative stress and/or oxidative injury (e.g., catalase, APX, GST, GAPDH, many molecular chaperones, and ALDH). In parallel, reduction of levels of proteins involved in electron transport and

unchanged levels for enzymes involved in carbon assimilation were determined in AOX lacking cells. Thus, AOX respiration was suggested as an important general mechanism of growth regulation in plant cells in response to nutrient availability and in maintaining cellular redox and carbon balance.

Recently, Alves et al. [32] reported on the analysis of *L. albus* leaf apoplastic proteins in response to boron (B) deficiency. Boron is mainly localized in the plant cell wall of vascular plants and some green algae, but its precise function is still not fully understood. On the protein level, 23 spots varied significantly between apoplastic fractions of control and B deficiency samples. Out of this, nine polypeptides could be identified by MS, all of them being involved in stress and plant defense responses, namely PR1-like protein, glucanase, class III chitinases, TLPs, and an expansin-like protein. However, *de novo* synthesis during B deficiency could be detected only for PR1-like protein, while the remaining proteins were also responsive to water deficiency.

Heavy Metal Stress. The widespread and undaunted use of chemicals in the past led to a worldwide accumulation of pollutants in our environment. Beside organic chemicals, heavy metals, especially Cd, As, and Pb, are the main pollutants found in groundwater and soil. Plants, as sessile organisms, are exposed to soluble metals to a high degree, and thus analyses of the responses to heavy metal stress and detoxification mechanisms are of strong interest.

In order to evaluate the responses of a forest tree species (*P. abies* L. Karst.) to heavy metals in aqueous solution, suspension cultured spruce cells were treated with different concentrations of CdSO₄ (50–500 µM), NaHAsO₄ (1.5–80 µM), or PbCl₂ (10–150 µM) [33]. Measurements of oxidative stress, antioxidants, and basic detoxification enzymes involved in plant defense reactions were performed, indicating a much broader reaction of plants to heavy metal stress than simple activation of oxidative protection and chelation. Nevertheless, as an early response, the onset of a strong oxidative burst was observed, expressed by accumulation of H₂O₂ accompanied by drastic changes in GSH levels. In parallel, strong induction of GST subunits was observed and the rise of a new GST isoform occurred. The most potent inducer of stress responses was Cd, followed by As and Pb in descending order of effectiveness. Using a signature peptide approach, changes on GST levels in response to copper, a promoter of oxidative stress, and benoxacor, an herbicide safener, were investigated [34]. Eight GSTs were identified in *Arabidopsis* seedlings; of these, one GST showed higher accumulation during benoxacor treatment, while copper treatment resulted in induced levels of an additional three GSTs. In contrast, one GST was significantly less abundant in response to both treatments. Recently, the early responses of *A. thaliana* cells to Cd exposure were investigated by a combined proteome and metabolite approach [35]. As a result, proteomic analysis revealed induced expression of a large set of proteins associated with many different cellular processes, such as energy production, cell rescue, cell defense, protein folding, and several metabolic pathways during cadmium exposure. For some of the candidate proteins the accumulation pattern in response to Cd exposure could be confirmed on the transcript level by Northern blot and RT-PCR analysis. Metabolite analysis by LC–ESI–MS

revealed the strongest variation on the metabolite level for six different families of phytochelatins, highlighting especially the activation of carbon, nitrogen, and sulfur metabolism, when taking together the results of the protein and metabolite profiling.

The specific involvement of peroxidases and other proteins in responses to excess Mn levels was investigated in leaf apoplasts from Cowpea (*V. unguiculata*) using 2-DGE of water-soluble proteins, with BN or IEF separation as a first dimension, combined to MS [36]. High concentration of free Mn in the apoplastic washing fluid and simultaneous induction of formation of brown spots and callose as well as activation of guaiacol- and NADH-POXs, indicated a particular role of free apoplastic Mn in mediating Mn-toxicity. In parallel, a significant increase in levels of proteins released to the apoplast was observed, among them PR proteins (e.g., glucanase, chitinase) and TLPs.

With the aim to identify proteins that may play a role in the exceptional degree of Ni tolerance of the hyperaccumulator plant *A. lesbiacum*, a 2-DGE approach was applied to roots and shoots from plants treated with either 0.3 mM NiSO₄ for 48 h or 0.03 mM NiSO₄ for 28 days [37]. When plants were grown for 28 days at 0.03 mM NiSO₄, a concentration still optimal for growth of this species and sufficient to lead to hyperaccumulation of Ni in the shoot, very few proteins were found to be changed in abundance. In contrast, 11 proteins with increased and one protein with decreased accumulation relative to control conditions were detected when plants were grown under 0.3 mM NiSO₄ for 48 h. Among the proteins strongly affected by short-term exposure, putatively involved in Ni tolerance and accumulation, were polypeptides involved in sulfur metabolism (e.g., OASTL, SHMT, and SAT), protection against ROS (e.g., LADH, a putative NADP-dependent oxidoreductase, GST, and mannose-6-phosphate reductase), or heat shock response (e.g., HSP101, and HSP70).

A promising and cost-effective method for metal decontamination of soil and water is *in situ* decontamination by use of crop plant species with high values of growth rate and biomass production, such as maize. Aiming to elucidate the potential of maize to react effectively to As exposure, a proteome approach was reported by reference 38. Maize seedlings were exposed to 300 μ M sodium arsenate or 250 μ M sodium arsenite for 24 h in a hydroponic system, and differentially displayed proteins were studied by 2-DGE. Out of 700 detected proteins, 20 were selected to be reproducible up- or down-regulated by the metalloid. Identification could be achieved for 11 of them, from which seven were highly involved in cellular homeostasis for redox perturbation (e.g., three SODs, two glutathione peroxidases, one peroxiredoxin, and one *p*-benzoquinone reductase), besides four additional, functionally heterogeneous proteins (e.g., ATP synthase, succinyl-CoA synthetase, cytochrome P450, and guanine nucleotide-binding protein β subunit). These results strongly indicated that the induction of oxidative stress is a main feature associated with As toxicity in plants. Later on, protein changes in maize plantlets in response to potassium dichromate treatment were analyzed [39]. Analysis of proteins differentially expressed during Cr exposure revealed, in principal, polypeptides involved in oxidative stress tolerance (e.g., SOD, 1-Cys peroxiredoxin/Rab 24, and glyoxylase I) and other stress defense pathways (e.g., Met synthetase, and SAM synthetase), besides proteins implicated in

sugar metabolism (e.g., *D*-GADPH, and ALDH) and energy production (e.g., ATP synthetase α chain).

Gaseous Pollutants. Among environmental gaseous pollutants, O₃, CO₂, and SO₂ are well known to cause serious affection of plant growth, development, and productivity. However, despite intensive studies of physiological and biochemical responses of plants to those pollutants, only a few number of studies have addressed the influence of these factors on the protein level. Only two important monocot species, namely wheat and rice, have been investigated for there responses to O₃ and SO₂ pollution using a proteomics approach. Results of those studies were summarized in reference 40. As major responses to physiological doses of O₃ and SO₂, changes in protein phosphorylation patterns were reported in parallel with differential expression of RuBisCO subunits and polypeptides involved in early stress responses (e.g., PR proteins, OsPR5 and OsPR10) as well as proteins related to oxidative stress reactions (e.g., APX and SOD).

In *A. thaliana* plants grown under elevated atmospheric CO₂ conditions, the changes in expression of soluble proteins and leaf metabolite levels were analyzed [41] using biochemical assays and 2-DGE. After long-term exposure, *Arabidopsis* plants used in this study displayed many of the common acclimation responses to elevated CO₂ conditions and sufficient nitrogen observed in previous studies, such as delayed flowering, accumulation of starch and sucrose, and acclimation of photosynthesis. In contrast, only a few differences were observed on the protein level when compared to control treatment. No protein appeared *de novo* or disappeared during CO₂ treatment, and significant differential expression was shown for 13 protein spots on at least one harvest date. Only six proteins with diverse function (e.g., myrosinase precursor, luminal BiP, NDPK, and OEC) were identified, allowing only preliminary conclusions on the cellular responses underlying the physiological changes following elevated CO₂ conditions.

Herbicide Stress. A number of efforts have recently been made in order to investigate the toxicity of frequently used herbicides on nontarget species and the subsequent plant responses. To this purpose, the proteome of wheat coleoptiles treated with fluxofenim, an herbicide safener which is known to protect cereal crops from herbicide injury by increasing the expression of herbicide-detoxifying proteins, such as GSTs, were investigated [42]. Twenty proteins were identified, 18 of which were shown to be inducible by the safener. Among them, 15 were identified as GST subunits, whereas the remaining three showed homology to the aldo/keto reductase family and to proteins involved in glycolysis and Krebs cycle. For elucidation of the entire herbicide detoxifying pathway in grapevines, a proteomics approach was applied on tissues subjected to herbicide stress [43]. Treatment with the systemically active herbicide flumioxazin resulted in impaired photosynthesis, stimulation of photorespiration, induction of the enzymatic antioxidant system, and a strong stimulation of key components of plant defense (PR10 proteins), reflected by differential expression patterns for 33 protein spots. More recently, Kelley et al. [44] reported on the evaluation of auxin-responsive genes in soybean which are specifically induced by plant growth regulator (PGR) herbicides. Expression of *GH3* was highly induced by PGR herbicides

at the RNA and protein level and was not affected during environmental stresses (heat, drought, and salt stress) or viral infection, indicating its potential for use as diagnostic marker for PGR herbicide injury.

Excessive or Inadequate Light. In contrast to the intensive study of the influence of water and nutrient status on plant proteomes, plant responses to light and temperature stress are rare. The UV-B and desiccation-tolerant terrestrial cyanobacterium *N. commune* was used to monitor proteome changes in the membrane, cytosolic, and the extracellular fractions during defined UV irradiation [45]. A large number of proteins with altered accumulation were observed, and responses to UV-B irradiation were divided into (a) an early shock response influencing 214 protein spots and (b) a late acclimation response involving 279 protein spots. The mainly transient shock responses included many membrane or membrane-associated proteins, while the acclimation response mainly included cytosolic proteins. Proteins mostly induced in the extracellular fraction were SOD and a water stress protein. Negative effects of UV-B irradiation on nitrogen uptake were evaluated in maize, indicated by decreased leaf and root NR activities and reduction of total nitrate content [46]. Results from this study indicated that the response to UV-B stress on NR activity relies on modulation at the post-transcriptional level rather than transcriptional regulation of NR isoforms.

In a study combining 2D-DIGE technology with MS and immunodetection responses to UV-B irradiation, maize leaves were analyzed in a flavonoid-deficient line (*b, pi* W23), along with two landraces from high altitudes (Cacahuacintle and Confitte Puneño) with improved UV-B tolerance [47]. Changed accumulation patterns were observed for a large number of proteins, which correlated to expression levels obtained earlier by microarray hybridization and real-time RT-PCR. A number of proteins differentially expressed during UV-B treatment in W23, and with higher constitutive levels under these conditions in cultivars from high altitudes, have been identified, making them potential candidate traits conferring UV-B tolerance. In addition, at least one protein (pyruvate phosphate dikinase) was found to be post-translationally regulated by reversible phosphorylation during UV-B exposure.

The response of *A. thaliana* to excessive light was gauged from experiments with isolated chloroplasts [48]. Both the soluble and insoluble proteins were subjected to 2-DGE, revealing 52 proteins that changed in abundance as a result of the imposed stress. Among these were a number of polypeptides involved in photosynthesis, as well as some known light-stress-related proteins, such as HSP, DHAR, and SOD. Lately, the high light responses of the soluble luminal proteome in *Arabidopsis* wild-type and the ascorbate-deficient mutant *vtc2-2* were investigated in a comparative proteomics study [49]. After 5 days of exposure to high light, both wild-type and *vtc2-2* plants accumulated anthocyanins, increased their total ascorbate content, and lost 10% of PS II efficiency. Restricted availability of reduced ascorbate (20% of wild-type) in the mutant resulted in repression of anthocyanin accumulation (34% of wild-type) and thus enhanced oxidative stress in *vtc2-2*. Proteome analysis revealed 45 protein spots with significantly changed expression levels as a consequence of genotype, light treatment, or both. Main genotypic differences were reduced levels of FeSOD, and enhanced

levels of Cu-ZnSOD in *vtc2-2* when compared to wild-type. Systematically higher levels were observed for a steroid dehydrogenase-like protein in *vtc2-2*, and during high light stress a stronger accumulation of PsbS was observed in wild-type plants. Results were also confirmed by Western blot analysis. Most significant responses to high light exposure were the strong accumulation of YCF37 likely involved in PS I assembly, a number of fibrillins, a flavin reductase-like protein, and an aldolase. All of these proteins are located on the thylakoid-associated plastoglobules, indicating the important role of this subcompartment in defending stress by high light intensities. In addition, protein profiling of plastoglobules from chloroplasts (*A. thaliana*) and chromoplasts (*C. annuum*) revealed differential accumulation of metabolic enzymes [50]. By using MS, it was shown that the chloroplast plastoglobules consist mainly of seven fibrillins and an additional 25 proteins likely involved in isoprenoid biosynthesis, lipid production, and carotenoid cleavage. Also, four unknown ABC1 kinases were identified. Reactions to high light stress and degreening were investigated using SIL and determined mainly differential responses to the treatment within the fibrillins family. From the results, metabolic function in breakdown of carotenoids and oxidative stress defense was derived for plastoglobules in chloroplasts, while plastoglobules in chromoplasts might also be active in carotenoid conversion, and thus a functional metabolic link between the inner envelope and thylakoid membranes was suggested.

High Temperature. To determine the effect of heat stress on the mature wheat grain proteome mainly represented by the non-prolamines fraction, a 2-DGE approach was used [51]. Out of 43 protein spots reproducibly responding to heat treatment, 24 were found to be up-regulated while 19 were down-regulated. Decreased expression levels were observed for proteins involved in starch synthesis, carbohydrate metabolism, and ATP production. Among the proteins induced by heat stress were HSPs as well as elongation and eukaryotic translation initiation factors. In accordance with these findings, the identification and characterization of a constitutively expressed stress-inducible smHSP (*AtHsp15.7*) from *A. thaliana*, was reported [52]. Expression studies by RT-PCR revealed strongly induced expression of *AtHsp15.7*, targeted to the matrix of plant peroxisomes, by heat and oxidative stress.

A proteome profiling of *Populus euphratica*, ecologically characterized as a pioneer because of surviving extreme temperatures as well as drought and salt stress, was performed to study the mechanisms leading to endurance of woody plants under heat stress [53]. Main effects of moderate heat stress on the leaf proteome were changes in the abundance of proteins involved in photosynthesis (long-term up-regulated) and carbon metabolism (long-term down-regulated). Short-term up-regulated proteins were related to membrane destabilization and cytoskeleton restructuring, sulfur assimilation, thiamine and hydrophobic amino acid biosynthesis, and protein stability. The only redox homeostasis-related enzyme found to be early long-term up-regulated in this study was thioredoxin *h*, highlighting its specific role in response to moderate heat exposure in woody plants.

Low Temperature. Low temperature has a detrimental effect on male reproductive development in rice, and 2-DGE profiling of the rice anther proteome (see

also Chapter 17) has been able to demonstrate the enhanced and induced selective partial degradation of proteins at the young microspore stage as a result of cold treatment [54]. Analysis of responses to progressively low temperature stress treatment, with a focus on leaves of rice seedlings, revealed 60 protein spots with induced expression pattern and different dynamic patterns when compared to control treatment [55]. Cold responsive proteins include factors of protein biosynthesis, protein degradation, and protein stabilization in addition to enzymes involved in cell wall biosynthesis, antioxidative/detoxifying enzymes, and proteins linked to energy pathway and signal transduction. Results from this study implicated that protein quality control mediated by chaperones and proteases as well as augmentation of cell wall components play important roles in tolerance of rice to cold stress. From the proteins found to be responsive to cold treatment, 43% were predicted to be located in the chloroplast, implying that the chloroplasts are one of the organelles mostly affected by cold stress. More recently, the rice cultivar Nipponbare was used for a comparative proteome approach to gain new insights into chilling stress responses [56]. By using MS analysis, 85 differentially expressed proteins were identified, mainly involved in photosynthesis (35%), energy metabolism (8.2%), photorespiration (7.1%), carbon metabolism (7.1%), redox homeostasis (5.9%), translation (5.9%), signal transduction (4.7%), protein processing (4.7%), sulfur (2.4%), and nitrogen metabolism (2.4%) as well as RNA processing (1.2%). Several of the proteins identified showed enhanced degradation during chilling stress, especially proteins involved in photosynthesis. Interestingly, gene expression analysis for 44 different proteins by quantitative real-time PCR revealed that the mRNA levels were not well-correlated to the protein levels. In addition, proteome analysis of the cold stress response in seedlings of the cold-sensitive rice cultivar Liangyoupeijiu was reported [57]. Among the proteins differentially expressed were mainly components of carbon metabolism (down-regulated), in addition to proteins involved in stress-resistance and cell structure damaging (up-regulated).

For identification of proteins associated with early cold acclimation in *Arabidopsis* leaves, proteome analysis of highly purified PM fractions was performed [58]. During the first day of cold acclimation, 38 proteins with altered expression levels were observed, mostly involved in membrane repair by membrane fusion, protection of the membrane against osmotic stress, enhancement of CO₂ fixation, and proteolysis. In parallel, the influence of cold stress on the nuclear proteome of *Arabidopsis* was investigated [59]. In the purified nuclear fraction, 184 spots corresponding to 158 different proteins could be identified. Out of these, 54 were up- or down-regulated more than two-fold during cold treatment. Six proteins were selected for further characterization by Northern blot analysis and transient expression studies. Data revealed confirmation of expression levels on the RNA level, and showed localization of two proteins in the cytoplasm and the nucleus while four proteins were detected exclusively in the nucleus. Recently, the cold stress responses of *A. thaliana* have been studied using DIGE analysis of the leaf proteome [60]. This identified the accumulation of a number of proteins—in particular, dehydrins and low-temperature-induced protein [78].

Effects of cold acclimation on protein composition and physiological parameters were also investigated using a forest tree species [61]. Three-month-old poplar

plants were exposed to chilling stress (4°C) and leaf material analyzed relative to control-treated material. Rapid accumulation of sucrose, glucose, fructose, and trehalose was observed under chilling conditions, while raffinose content increased after one week of treatment. Protein analysis revealed 30 spots with altered expression patterns, and by MS about one-third of spots accumulating during cold stress were identified as chaperone-like proteins. In addition, dehydrins and other late embryogenesis abundant proteins were also activated or newly synthesized. Down-regulation was observed for proteins involved in cell wall synthesis and energy production. Later on, Western blot analysis revealed induced expression of three families of stress-related proteins—that is, dehydrins, stress protein 1, and HSP70—by low temperatures, thus complementing the results obtained earlier [62].

Common Changes in the Proteome in Response to Different Abiotic Stresses

Although we have summarized the recent proteome studies on stress responses for single stress factors, it is evident that under many natural conditions the plants have to respond to a combination of stresses. Consistent with biochemical and molecular data, proteome approaches also indicate an overlap in the cellular responses to different stresses, such as detoxification of ROS by enzyme components as a ubiquitous element of defense. It should also be mentioned that due to limitations in current techniques, proteome analysis is likely to be restricted to the more abundant proteins of central metabolic pathways, implying that specific aspects of cellular responses characteristic to certain stress factors might not yet be detected with current methodology. However, a survey of published data indicates that in many experimental systems, fundamental reprogramming of central metabolic routes can be monitored. Whereas changes in the abundance of proteins representing house-keeping functions are covered by current proteome methods, analysis of signaling events in response to abiotic stresses is less evident. Targeted analysis of changes in the phosphoproteome of rice seedlings upon treatment with stresses or hormones revealed phosphorylation of enzymes, but also identified Ca^{2+} -signaling processes [63]. NO signaling was identified as a major route in the cellular responses after pathogen attack, but is also part of physiological processes such as stomatal closure, growth, and development. Investigation of *S*-nitrosylation in *Arabidopsis* identified a range of candidates representing metabolic as well as signaling proteins [64]. Clearly, more of such studies are needed to reveal common and specific cellular responses to abiotic stresses.

In a comparative analysis to define the cellular mechanisms operating in plant trichomes, the soluble proteins of whole leaves and isolated trichomes from tobacco were subjected to 2-DGE [65]. Proteins related to stress defense responses (cytosolic Cu-ZnSOD, glutathione peroxidase, PR-2, PRQ, and an *A. thaliana* stress-inducible protein) were strongly represented among the products specifically enriched in tobacco trichomes. The specificity of expression of these proteins in the trichomes was confirmed by protein gel blotting and enzyme assays, indicating that tobacco trichomes constitute a valuable system to study defense response networks on the proteome level.

When investigating rice seedlings, the expression of SOD was a common response to cold, drought, salt, and ABA treatment while the expression of three proteins with unknown function was enhanced by salt and ABA stresses [13].

In a study combining determination of physiological parameters of pea leaves—such as photosynthetic and respiratory rates as well as lipid peroxidation product accumulation—with BN and 2-DGE separation of isolated mitochondrial proteins, the differential impact of environmental stresses on the pea mitochondrial proteome was investigated. Data suggested that herbicide treatment exerted a more severe stress on the mitochondria than did chilling and drought. Differential induction of HSPs and specific changes in other proteins suggested divergence of cellular responses upon the various stresses [66].

41.3 COMPARISON OF TRANSCRIPTOMICS AND PROTEOMICS DATA IN ANALYSIS OF PLANT DEFENSE RESPONSES AGAINST ABIOTIC STRESS FACTORS

The current status of analytic depth is, in general, considerably more advanced for transcript profiling than for proteome analysis. Chemically, nucleic acids are more uniform than proteins, rendering the comprehensive analysis of the proteome more challenging. For model organisms and crop species like *Arabidopsis*, rice, barley, potato, and maize, expression for a large fraction of all genes can be simultaneously analyzed. Yet the analysis of the proteome is mandatory, because quite a number of studies have shown that there is no strict correlation between transcript levels and protein abundance, such as in yeast [67], mouse [68], or *Arabidopsis* [69]. Lack of correlation for levels of certain transcripts and their encoded proteins at certain experimental time points might in many instances be related to different kinetics, while in other cases might point to independent regulatory events for transcripts and resulting proteins. Even important as an argument for proteome approaches, expensive and time-consuming in many cases, are the many possibilities for PTMs of proteins with fundamental impact on function not amenable by transcript profiling approaches. In the following, before dealing with a few studies pioneering in combined comprehensive transcript and protein profiling (summarized in Table 41.2), we will first summarize a number of selected studies on abiotic stress responses based solely on transcript profiling.

Many aspects of the transcriptional regulatory network in response to dehydration and cold stress in *Arabidopsis* plants have been elucidated by the pioneering work of the group of Yamaguchi-Shinozaki and Shinozaki [70]. Expression of many genes is influenced by abiotic stresses, among them a number of transcription factors that participate in the signaling cascades in response to drought or cold. ABA has been identified as a major signal metabolite under drought and salinity. Exogenous application of ABA also promotes the expression of a range of genes also responsive to drought and cold stress, although the role of this phytohormone in cold stress signaling is unclear. On the other side, quite a number of genes responsive to drought or cold stress do not respond to exogenous application of ABA [70, 71]. These results

TABLE 41.2. Reports on Transcriptome/Proteome Analyses of Cellular Responses to Different Abiotic Stress Factors

Plant Species	Stress Conditions	Investigated Material	Reference
<i>Arabidopsis thaliana</i>	Drought and cold		70
	Drought and salinity		71
	Drought, cold, and salinity	Plantlets	72
	Cold, salinity, and osmotic stress	Leaves and roots	73, 75
	Cold	Leaves	74
	Sulfur deficiency	Seedlings, leaves, and roots	80 (review)
<i>Oryza sativa</i> L.	Salinity	Leaves and roots	76
<i>Zea mays</i> L.	Uv-B	Leaves and roots	78
	Cold and desiccation	Embryo axes of seeds	81
<i>Prunus persica</i> L.	Cold and photoperiod	Bark	79

are summarized in current models of signal transduction to osmotic and cold stress; these models define ABA-dependent and ABA-independent pathways and overlaps in the signaling cascades between both stress factors [70]. When analyzing the expression profiles of 7000 *Arabidopsis* genes in response to various abiotic stresses using full-length cDNA microarray, the transcripts of 53, 277, and 197 genes were five-fold increased by drought, cold, and high salinity, respectively, compared to controls [72]. Twenty-two stress-inducible genes responded to all stresses; and among the genes stimulated by either of the stresses, about 11% were TFs. In a similar study on salt, osmotic, and cold stress responses in *Arabidopsis*, Kreps [73] expression of more than 2400 out of 8100 genes was modified twofold by any of these stresses, indicating that overall gene expression is influenced considerably. Roots and leaves showed different changes. The largest induction upon stress treatments was observed for *LTI/COR78/RD29A* (At5g52310). The promoter of this gene contains the conserved DRE sequence also found in the promoter regions of many other drought- and cold-inducible genes. Related *cis*-acting elements (CRT) and LTRE also regulate cold-inducible promoters.

The cold acclimation response of *Arabidopsis* wild-type was studied by monitoring the expression of 8000 genes and identified 218 genes increasing and 88 genes decreasing threefold or more at one of the timepoints. Forty-eight of these genes were TFs. This study also included analysis of transgenic lines overexpressing CBF proteins acting as transcriptional activators binding to CRT/DRE elements, allowing one to define the regulon of CBF. Results indicate that only 12% of the cold-responsive genes were members of the CBF regulon, and at least 28% of the cold-responsive genes were not regulated by the CBF TFs, indicating the occurrence of other low-temperature regulons. Plants with ectopic CBF expression grown under warm temperature showed down-regulation of eight genes otherwise repressed by low temperatures, demonstrating the equal importance of up-regulation and down-regulation of genes in response

to cold stress [74]. The studies outlined so far clearly show the significance of TFs in establishing regulatory networks operative particularly in early stress defense responses consistent with a study evaluating the expression profiles of 402 TFs representing roughly one-third of all *Arabidopsis* TFs in response to environmental stresses [75].

The gene expression profiles during the early phase of salt stress were investigated in a salt-tolerant rice line with microarrays including 1728 cDNAs from libraries of salt-stressed roots [76]. NaCl at 150 mM reduced photosynthesis considerably within minutes, and kinetics of transcript levels were monitored from 15 min up to one week. Fifteen percent of the transcripts were up- or down-regulated within 1 h. The differences between control and stress plants became less pronounced in the later stages consistent with an adaptation to the stress. The up-regulated functions such as protein synthesis, protein turnover, and defense showed different kinetic patterns. Comparative analysis of a salt-sensitive line demonstrated a delay of the initial responses of the salt-tolerant line, with a later decline of transcript levels followed by cell death [76].

Identified transcript factors of the stress defense signaling network have been used for overexpression in different plant species, and the generated lines were tested for improved stress tolerance (see Table 1 in reference 77) demonstrating the capacity to improve stress-related traits in crops. Responses in the transcript levels upon UV-B stress treatment were followed in maize using the Unigene I array containing 5664 ESTs [78] and led to the identification of a number of UV-responsive genes.

Global analysis of gene regulation by low temperature was also addressed in peach trees by using suppression subtractive hybridization libraries as a tool to identify relevant candidates. Low temperature resulted in the up-regulation of a number cold-responsive and stress-related genes and the down-regulation of genes with house-keeping function [79]. A number of studies have been performed on the elucidation of plant adaptive mechanisms triggered by sulfur deficiency on the transcript level (reviewed in reference 80). Based on the summarized results, a response network linking several pathways related to JA, oxidative stress response, auxin, and flavonoid to sulfur metabolism was suggested. Further post-genomics studies, including analysis of gene products and metabolites, are needed to define molecular mechanism underlying the responses to limited sulfur nutrition.

A failing of much of the literature is the lack of supportive functional data from genetic and other approaches, and projects combining comprehensive protein expression mapping with transcript expression profiling are still uncommon. An example of this combined approach has been reported in reciprocal maize hybrids differing in their reaction to cold germination and desiccation tolerance [81]. Most of the differentially expressed proteins and/or mRNAs were implicated in more than one structural, enzymatic, or regulatory function and could be localized in various sub-cellular fractions. In further studies, pleiotropic effects should be eliminated, leading to a set of traits closely involved in the control of cellular processes and resulting in the manifestation of differences in cold germination and desiccation tolerance phenotypes.

41.4 CONCLUSIONS

Proteomics analysis of plant responses to different abiotic stresses has gained substantial interest within recent years. Cellular responses to all common stress factors have been studied in a variety of plant systems. There is a clear focus on drought, cold, and salinity stress, which have the most serious impact on agronomic production, consistent with other biochemical and molecular work performed on these stresses. Collectively, the publications clearly demonstrate that proteome approaches can provide a global analysis of the responses of plant systems toward abiotic stresses. Quite a few protein spots being up- or down-regulated subsequent to stress treatments have been described in earlier targeted analyses, serving as kind of a positive control to validate the proteome analysis. In comparison to current transcript analysis, the proteome studies are less comprehensive, due to limitations in the separation of complex protein extracts. Also, many transcript profiling studies explore the early events in response to stress exposure including the analysis of signaling cascades, whereas most current proteome approaches reflect the final down-stream responses. Analysis of signaling events on the protein level, along with the further investigation of PTMs, is a priority for future proteome analysis of abiotic stress responses in plants.

41.5 FIVE-YEAR VIEWPOINT

Within the next five years, depth of the proteome analysis for the characterization of stress responses will be substantially refined to cover a wider fraction of the protein complement. Different strategies can be applied for this purpose. A few papers already describe the analysis of organelles from stressed plants, and extension of this work will enable to enhance resolution of the proteome studies. In addition to that, separate analysis of prefractionated protein extracts will allow a higher resolution of the proteomic analysis. Improved proteome methods such as label-free quantitative LC-based techniques will also contribute to higher analytical depth. A major strength of the proteome approaches lies in the capability to monitor PTMs, such as phosphorylation among many other possibilities. These PTMs represent fundamental aspects of stress defense mechanisms, and improved understanding of these events will be of considerable importance in further defining signaling cascades and regulatory circuits operating in stressed plants. Overall analysis of stress defense responses in plants with the aim to generate more tolerant crops will greatly benefit from a combination of techniques applied to the one and same material. For many plant species, sufficient resources and tools have become available to perform transcript and protein profiling in combination. Even important the comparative analysis of genotypes and related mapping populations contrasting in stress tolerance will narrow the number of putative candidate genes. This combination of valuable genetic resources with recent “omics” techniques for comprehensive analysis of transcripts, proteins, and metabolites has been termed genetical genomics [82] and will contribute to the progress in functional genomics, including characterization of stress defense networks. In addition, strong bioinformatic tools (e.g., pathway visualization) will be mandatory for

these combined approaches. Proteomic studies within the next five years will also be beneficial for the functional characterization of previously identified candidate genes, by analyzing mutants and transgenic lines to study effects of modified target gene expression and thereby defining the metabolic or regulatory context of the gene product.

REFERENCES

1. Bray EA, Bailey-Serres J, and Weretilnyk E. 2000. Responses to abiotic stresses, In: *Biochemistry and Molecular Biology of Plants* (Buchanan B, Gruissem W, and Jones R, eds.), American Society of Plant Physiologists, Rockville, MD, pp. 1158–1203.
2. Ashraf M and Harris PJC. 2005. *Abiotic Stresses: Plant Resistance Through Breeding and Molecular Approaches*. Haworth Press, Binghamton, NY.
3. Inze D and Van Montague M. 2002. *Oxidative Stress in Plants*. Taylor & Francis, London.
4. Riccardi F, et al. 1998. *Plant Physiol* **117**:1253–1263.
5. Vincent D, et al. 2005. *Plant Physiol* **137**:949–960.
6. Pinheiro C, Kehr J, and Ricardo CP. 2005. *Planta* **221**:716–728.
7. Hajheidari M, et al. 2005. *Proteomics* **5**:950–960.
8. Salekdeh GH, et al. 2002. *Proteomics* **2**:1131–1145.
9. Riccardi F, et al. 2004. *Plant Physiol Biochem* **42**:1003–1011.
10. Rakwal R and Komatsu S. 2004. *Mol Biol Rep* **31**:217–230.
11. Ali GM and Komatsu S. 2006. *J Proteome Res* **5**:396–403.
12. Jorge I, et al. 2006. *Proteomics* **6**:S207–S214.
13. Abbasi FM and Komatsu S. 2004. *Proteomics* **4**:2072–2081.
14. Yan SP, et al. 2005. *Proteomics* **5**:235–244.
15. Kim DW, et al. 2005. *Electrophoresis* **26**:4521–4539.
16. Parker R, et al. 2006. *J Exp Bot* **57**:1109–1118.
17. Zörb C, et al. 2004. *Plant Sci* **167**:91–100.
18. Askari H, et al. 2006. *Proteomics* **6**:2542–2554.
19. Kav NNV, et al. 2004. *Ann Appl Biol* **145**:217–230.
20. Lee S, et al. 2004. *Plant Cell* **16**:1378–1391.
21. Dani V, et al. 2005. *Proteomics* **5**:737–745.
22. Fulda S, et al. 2000. *Eur J Biochem* **267**:5900–5907.
23. Huang F, et al. 2006. *Proteomics* **6**:910–920.
24. Ndimba BK, et al. 2005. *Proteomics* **5**:4185–4196.
25. Zhang XX, Takano T, and Liu SK. 2006. *J Exp Bot* **57**:193–200.
26. Jackson MB and Colmer TD 2005. *Annals Bot* **96**:501–505.
27. Agarwal S and Grover A. 2006. *Crit Rev Plant Sci* **25**:1–21.
28. Kang JG, et al. 2004. *Proteomics* **4**:3549–3559.
29. Sarry JE, et al. 2006. *Electrophoresis* **27**:495–507.
30. Bahrman N, et al. 2004. *Proteomics* **4**:709–719.
31. Sieger SM, et al. 2005. *J Exp Bot* **56**:1499–1515.

32. Alves M, et al. 2006. *Plant Soil* **279**:1–11.
33. Schröder P, et al. 2003. *Environ Sci Pollut Res* **10**:225–234.
34. Smith AP, et al. 2004. *J Biol Chem* **279**:26098–26104.
35. Sarry JE, et al. 2006. *Proteomics* **6**:2180–2198.
36. Fecht-Christoffers MM, et al. 2003. *Plant Physiol* **133**:1935–1946.
37. Ingle RA, Smith JAC, and Sweetlove LJ. 2005. *BioMetals* **18**:627–641.
38. Requejo R and Tena M. 2005. *Phytochemistry* **66**:1519–1528.
39. Labra M, et al. 2006. *Chemosphere* **62**:1234–1244.
40. Agrawal GK and Rakwal R. 2006. *Mass Spectrom Rev* **25**:1–53.
41. Bae HH and Sicher R. 2004. *Field Crops Res* **90**:61–73.
42. Zhang O and Riechers DE. 2004. *Proteomics* **4**:2058–2071.
43. Castro AJ, et al. 2005. *J Exp Bot* **56**:2783–2795.
44. Kelley KB, et al. 2006. *Weed Sci* **54**:220–229.
45. Ehling-Schulz M, et al. 2002. *Mol Microbiol* **46**:827–843.
46. Quaggiotti S, et al. 2004. *Plant Sci* **167**:107–116.
47. Casati P, et al. 2005. *Mol Cell Proteomics* **4**:1673–1685.
48. Phee BK, et al. 2004. *Proteomics* **4**:3560–3568.
49. Giacomelli L, Rudella A, and van Wijk KJ. 2006. *Plant Physiol* **141**:685–701.
50. Ytterberg AJ, Peltier JB, and van Wijk KJ. 2006. *Plant Physiol* **140**:984–997.
51. Majoul T, et al. 2004. *Proteomics* **4**:505–513.
52. Ma CL, et al. 2006. *Plant Physiol* **141**:47–60.
53. Ferreira S, et al. 2006. *Annals Bot* **98**:361–377.
54. Imin N, et al. 2004. *Proteomics* **4**:1873–1882.
55. Cui SX, et al. 2005. *Proteomics* **5**:3162–3172.
56. Yan SP, et al. 2006. *Mol Cell Proteomics* **5**:484–496.
57. Yang PF, et al. 2006. *J Integrative Biol* **48**:945–951.
58. Kawamura Y and Uemura M. 2003. *Plant J* **36**:141–154.
59. Bae MS, et al. 2003. *Plant J* **36**:652–663.
60. Amme S, et al. 2006. *J Exp Bot* **57**:1537–1546.
61. Renaut J, et al. 2004. *Plant Biol* **6**:81–90.
62. Renaut J, Hoffmann L, and Hausman JF. 2005. *Physiol Plant* **125**:82–94.
63. Khan M, Takasaki H, and Komatsu S. 2005. *J Proteome Res* **4**:1592–1599.
64. Lindermayr C, Saalbach G, and Durner J. 2005. *Plant Physiol* **137**:921–930.
65. Amme S, et al. 2005. *Proteomics* **5**:2508–2518.
66. Taylor NL, et al. 2005. *Mol Cell Proteomics* **4**:1122–1133.
67. Griffin TJ, et al. 2002. *Mol Cell Proteomics* **1**:323–333.
68. Mijalski T, et al. 2005. *Proc Natl Acad Sci USA* **102**:8621–8626.
69. Mooney BP, et al. 2006. *Physiol Plant* **128**:237–250.
70. Yamaguchi-Shinozaki K and Shinozaki K. 2006. *Annu Rev Plant Biol* **57**:781–803.
71. Zhu J. 2002. *Annual Rev Plant Biol* **53**:247–273.
72. Seki M, et al. 2002. *Plant J* **31**:279–292.
73. Kreps JA, et al. 2002. *Plant Physiol* **130**:2129–2141.

74. Fowler S and Thomashow MF. 2002. *Plant Cell*. **14**:1675–1690.
75. Chen WQ, et al. 2002. *Plant Cell* **14**:559–574.
76. Kawasaki S, et al. 2001. *Plant Cell* **13**:889–905.
77. Zhang JZ, Creelman RA, and Zhu JK. 2004. *Plant Physiol* **135**:615–621.
78. Casati P and Walbot V. 2004. *Genome Biol* **5**:R16.
79. Bassett CL, et al. 2006. *J Am Soc Hortic Sci* **131**:551–563.
80. Hirai MY and Saito K. 2004. *J Exp Bot* **55**:1871–1879.
81. Kolipara KP, et al. 2002. *Plant Physiol* **129**:974–992.
82. Jansen RC and Nap JP. 2001. *Trends Gen* **17**:388–391.
83. Dubey H, et al. 2003. *Curr Sci* **84**:83–89.

PROTEOMICS OF BIOTROPHIC PLANT–MICROBE INTERACTIONS: SYMBIOSES LEAD THE MARCH

Ghislaine Recorbet and Eliane Dumas-Gaudot

42.1 INTRODUCTION

Both subterranean and aerial plant organs are constantly exposed to a multitude of microorganisms including viroids, viruses, bacteria, fungi, and oomycetes. PMIs exhibit an extraordinary wide range in their degree of dependence and outcome. It is largely admitted that most microorganisms fail to infect most plants because of pre-formed barriers or induction of defense responses following their recognition by the plant. However, some microbes have evolved strategies to overcome plant defenses and manage to infect plant tissues, resulting in plant–microbe compatible interactions [1]. In such associations, the mechanisms by which microorganisms succeed in attaining plant-derived nutrients encompass the destruction of plant cells to feed on the cell content (necrotrophic interactions), intercellular growth to feed on the apoplast (endophytic parasitic interactions), and the intracellular colonization of living plant cells (biotrophic interactions) [2].

Strikingly, the intracellular accommodation of a microorganism by a living plant cell can either lead to plant beneficial or detrimental effects, so that plant–biotroph interactions play major roles in agriculture [3, 4]. The distinction between parasitic

and beneficial biotrophic associations has its origin in the availability of nutrients to the microorganism. Typically, most parasitic biotrophs responsible for devastating plant pathologies like mildews, rusts, and smuts derive nutrients from shoot tissue without having any alternative energy source [5]. By contrast, most beneficial biotrophic microorganisms like mycorrhizal fungi and nitrogen-fixing rhizobia colonize root tissues and have access to nutrients outside the plant, raising the possibility for bidirectional nutrient movement and the development of a mutualist rather than a parasitic interaction [6].

Although leading to different outcomes, mutualist and parasitic biotrophs share the ability to penetrate the plant cell wall without causing the death of the host cell. This requires the differentiation of specialized intracellular accommodation structures corresponding to haustoria in most pathogenic interactions, haustoria termed arbuscules in the AM symbiosis, or bacteroids in the rhizobial nitrogen-fixing (RNF) symbiosis [3]. Because these structures are always surrounded by a host-derived plant PM, microbial biotrophs remain separated from the host cytoplasm. A resulting common feature of plant–biotroph associations corresponds to the existence of an interface, thought to be the main site of nutrient and signal flow between cells, that comprised the plant and the microbial membranes separated by a plant-derived apoplast [3, 6–8].

Along with these morphological similarities, the identification of structural, developmental, molecular, and/or genetic commonalities between different biotrophic PMIs has generated the conceptual idea that there should exist in plants some core compatibility modules conserved between parasitism and mutualism [2, 3, 7, 9]. Because the AM symbiosis is believed to be one of the oldest biotrophic interactions, it has been hypothesized that evolutionary younger microbial biotrophs might have recruited some aspects of the mutualistic program developed by plants to host AM fungi.

To establish a compatible interaction, biotroph microorganisms have to either avoid or suppress plant defense reactions together with redirecting the host metabolic flow to their benefit without killing the host. The mechanisms by which this is achieved are almost unknown, but proteins happen to take the lion's share in the paradigms that currently govern such PMIs by playing key roles in mediating recognition, signaling, nutrient transport, plant cell differentiation, and compatibility [5, 10]. In this respect, proteomics is likely to be one of the best methodologies to decipher some key features of biotrophy. However, partly because the analysis of the mechanisms governing plant disease susceptibility is only emerging compared to plant disease resistance [4], the vast majority of the proteomics studies performed so far to investigate biotrophic plant–microbe associations relate to AM and RNF symbioses [11–13], with only two recent reports concerning parasitic interactions with the fungal biotrophs *E. pisi* (powdery mildew) and *P. tritici-na* (leaf rust) [14, 15].

The usefulness of proteomics to gain knowledge about rhizospheric endosymbioses was first reviewed in 2004 [16]. Since that time, many additional symbiosis-related proteins (SMRPs) have been elucidated through proteomics approaches more specifically directed toward well-defined stages/tissues of root endosymbioses and/or biotrophy-related subcellular fractions. This chapter aims to provide an update of the contribution of proteomics to identify SMRPs during the last two years and is divided into three parts. In Section 42.2, key features making AM and RNF symbioses not

equally amenable to proteomic investigations are highlighted to underline that specific strategies are required to fulfill the identification of SMRPs. Sections 42.3 and 42.4 successively address the most recent proteomics data obtained when studying plant roots interacting with AM fungi and rhizobia. When possible, emphasis also is put on the putative existence of “biotrophy-related” proteins that show overlapping patterns of accumulation between taxonomically distant plant–biotroph interactions, either parasitic or mutualistic.

42.2 AM AND RNF SYMBIOSES: NO EQUAL FOOTING FOR PROTEOMICS

Differences in Developmental Features

AM symbiosis is the most widely distributed association, because it occurs between more than 80% of vascular land plants and about 150 soil fungal species belonging to the Glomeromycota [17]. AM fungi promote host plant growth by supplying phosphorus and other mineral nutrients absorbed from the soil solution. In turn, AM fungi are supplied with the organic carbon forms essential for them to achieve their full life cycle [7]. As a result, AM fungi are obligate plant biotroph microorganisms. The successful development of AM symbiosis is dependent on significant morphological and physiological modifications affecting both partners in a time-orchestrated way. Following spore germination and typical hyphal branching in response to plant exudates, the fungus differentiates appressoria at the root surface. Once reaching the inner root cortex, some hyphal branches penetrate the cortical cell walls and start to differentiate into highly branched structures called arbuscules. Concomitant with the fungal penetration into plant cells, the host PM invaginates along the penetrating hyphae to surround the arbuscule, increasing its surface area four to tenfold and resulting in the formation of the periarbuscular membrane. A new compartment composed of the membranes of both partners separated by apoplastic material is thus created [18] (Figure 42.1). This symbiosis is an extremely ancient PMI that has survived more than 400 million years on earth. AM fungi have been identified in fossils of early Devonian plants and are believed to have assisted them during the colonization of land [19].

The nitrogen-fixing symbiosis that forms between legume roots and soil bacteria that belong to the class Rhizobiaceae (called rhizobia) have evolved more recently, some 60 million years ago [20]. Once again, the plant provides its microsymbiont with photosynthetate in exchange for fixed nitrogen in the form of ammonium and amino acids. The interaction involves host–microorganism specificity via a recognition mechanism based on chitooligosaccharide molecules produced by the bacteria and called Nod factors [21]. After attachment to the host root hairs, bacteria colonize roots within infection threads and induce a meristematic activity in inner cortical cells, leading to the formation of nodule primordia. When released from the infection thread, the invading bacteria are packaged within plant membrane vesicles called peribacteroid (or symbiosome) membranes and further differentiate into bacteroids that fix nitrogen [22] (Figure 42.2). As root-enclosed rhizobia continue to divide, the

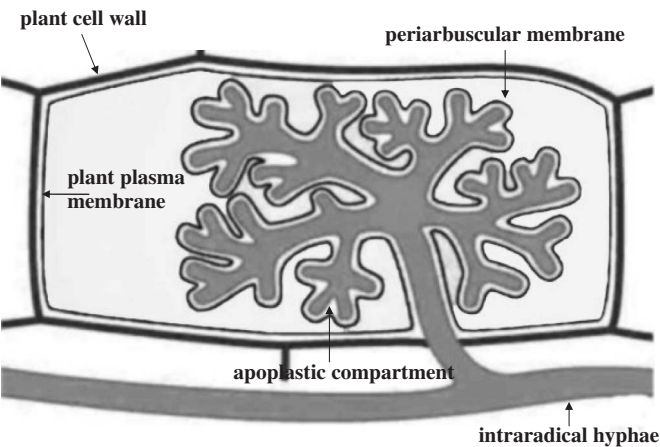


FIGURE 42.1. Schematic representation of an arbuscule-containing plant cell.

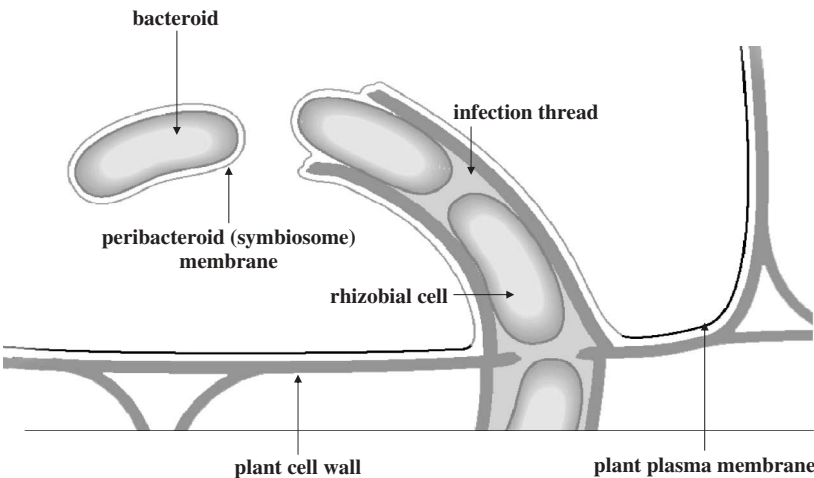


FIGURE 42.2. Schematic representation of an infection thread and bacteroid within a root cell nodule.

peribacteroid membrane may either divide with each bacterium or enlarge to contain up to eight bacteria [23].

On the whole, roots interacting with rhizobia and AM fungi are not equally amenable to proteomics analyses by reason of major differences in the available amount of pre-symbiotic and symbiotic materials. Free-living rhizobia can be grown axenically, which enables time-point inoculation-based experiments. On the contrary,

due to the obligate biotrophic status of the mycorrhizal microsymbiont, the fungal asymbiotic phase is restricted to spore germination and production of a limited amount of mycelium [24]. Likewise, infective rhizobial structures consisting of bacteria attached to legume root hairs can be isolated for further analyses [25]. This is not the case for fungal appressoria that first contact with plant tissues before root penetration. Additionally, studies during initial contact between AM fungi and plant roots suffer from difficulties in synchronizing development in the two symbionts [26]. Finally, RNF symbiosis culminates with the formation of a new organ (the nodule) that develops abundantly at the root surface. Symbiosomes also are large organelles of specific density and can easily be isolated from the nodule cells [27]. Additionally, the amplification of bacterial numbers inside plant membranes during the symbiotic process also makes rhizobia able to be collected by centrifugation techniques in sufficient quantity [22]. On the contrary, mycorrhizal symbiosomes only represent a limited amount of biological material that develop within host roots. Contrasting with haustoria of parasitic rusts and powdery mildews that can be purified from infected plant tissues [4, and references therein], fungal arbuscules yet have not been isolated in sufficient amount from root tissues.

Differences in Genomic Resources

Protein identification using MS techniques largely depends on genome sequence information. According to He et al. [28], a simple PMF is sufficient to identify 50–90% of proteins correctly when a complete genome sequence exists. For organisms lacking a complete sequence, MS/MS must be used, which reduces throughput. *In silico* MS-based protein identification relies on exact matching of masses of peptides and/or peptides fragments to corresponding masses calculated from sequences from databases entries and will fail to identify proteins when no sequence is present. Fortunately, many proteins in different species are homologous, permitting identification by their sequence similarity to known homologues from phylogenetically related species [29].

Contrasting with mycorrhizal endosymbiosis, nodulation is a highly host-specific interaction. Historically, on the basis of their particular suitability for structural and functional genomics, *L. japonicus* and *M. truncatula* have been chosen as model organisms for the analysis of specific aspects of legume biology that cannot be achieved with *A. thaliana*, including their unique ability to enter both AM and RNF symbioses [30]. Sequencing programs for these two genomes have been launched together with international EST projects. For now, more than 225,000 and 150,000 EST sequences from *M. truncatula* and *L. japonicus* have been deposited in GenBank dbEST (<http://www.ncbi.nlm.nih.gov/dbEST>), respectively, and large web-based genomic resources for legumes also are available [31]. Although AM symbiosis proteomics has mainly focused on AM fungi interacting with *M. truncatula* [reviewed in reference 16], the recent completion of genome sequences from the model crop *O. sativa* and the model tree *P. trichocarpa* will undoubtedly contribute to increase the identification of AM SMRPs [32, 33].

Regarding microsymbionts proteomics of RNF, symbiosis beneficiaries from availability of the complete genome sequences for many plant-colonising bacteria [34].

Among rhizobia, the genome of *M. loti* was the first to be sequenced [35]. Genome sequences from *S. meliloti* 1021, *B. japonicum* USDA 110, *R. etli* CFN42, and *R. leguminosarum* bv. *viciae* 3841 also are now completed. By contrast, very few genomes of plant-colonizing fungi so far have been annotated (<http://www.broad.mit.edu/annotation/cgi/>; <http://genome.jgi-psf.org/euk/>), and they mainly encompass those from pathogenic fungi including *B. cinerea*, *F. graminearum*, *M. grisea*, *U. maydis*, *S. sclerotiorum*, *S. nodorum*, *M. graminicola*, and *N. haematococca*. When regarding mutualistic fungi interacting with plants, complete genome sequence only is available for the ectomycorrhizal fungus *L. bicolor* that enters symbiosis with many northern temperate forest trees (<http://genome.jgi-psf.org/euk/>). Although the United States Department of Energy Joint Institute (<http://genome.jgi-psf.org/euk/>) also launched in 2004 a sequencing program for the genome of the AM fungus *G. intraradices*, no complete data are available at present. *G. intraradices* has been chosen as a model species for the AM fungi because of its relative small genome (15 Mb), its ubiquity in different ecosystems along with its ability to grow *in vitro* in dual culture with transformed carrot roots [reviewed in reference 33]. Several EST libraries have been constructed using different species and genera of AM fungi, and more than 5000 EST sequences have been deposited in GenBank dbEST (<http://www.ncbi.nlm.nih.gov/dbEST>). Overall, lack of sequence data corresponding to Glomeromycota in database still makes challenging the MS/MS-based identification of AM fungal proteins [36].

42.3 A 2005–2006 UPDATE OF THE CONTRIBUTION OF PROTEOMICS TO AM SYMBIOSIS

Since the pioneering work of Bestel-Corre et al. [37] in which the first MS-based identification of mycorrhiza-related proteins was reported, much effort has been input in methodological devices allowing enrichment of biological extracts for rare proteins relevant to the establishment and functioning of AM symbiosis. Depending on the symbiotic stage targeted and on the abundance of mycorrhizal material, different proteomic strategies have been set up. In mature mycorrhiza, subcellular proteomic approaches have been developed in the model legume *M. truncatula* to target symbiosis-related membrane proteins eligible for involvement in nutrient transport and signaling between symbionts upon arbuscule formation. Concomitantly, modification of the *M. truncatula* root proteome during the early stage of AM symbiosis also was investigated by comparing the protein patterns obtained from noninoculated roots and roots synchronized for appressorium.

Proteomics in Mature Mycorrhiza

In view of the large surface area in which both plant and fungal membranes are in close contact in arbuscule-containing cells, it is widely admitted that plant–fungus interfaces play a major role in nutritional and signal exchanges between both partners, leading to a considerable interest into characterizing the molecules transferred and the mechanisms involved. Subcellular proteomics, defined as the large-scale analysis

of proteins purified from a cell compartment, has emerged as a promising tool to enrich extracts with specific proteins. Regarding mycorrhizal associations, two main approaches have been conducted as an attempt to detect PM proteins of plant and fungal origin regulated upon AM symbiosis.

As a preliminary step toward the identification of AM-responsive root PM proteins, a procedure aimed at gaining access to proteins associated with total membrane structures (i.e., microsomes) was developed by combining a subcellular fractionation process to 2-DGE. By using a differential solubilization of membrane proteins in a C/M mixture, both peripheral and partially hydrophobic proteins were precipitated in the insoluble C/M fraction. Because strictly hydrophobic proteins escape 2-DGE-based proteomics, only this latter fraction could be analyzed by 2-DGE. Benefiting from the development of MS and bioinformatic tools, 96 out of the 440 well-resolved microsomal proteins could be identified by MALDI-TOF PMF when using *M. truncatula* clustered EST database for queries [38]. Among them, 21 proteins displayed one possible TMD. This procedure was also carried out to investigate the membrane-associated proteins that were regulated in response to the fully established AM symbiosis between *M. truncatula* and *G. intraradices* [39]. Comparison of 2-DGE profiles of the insoluble C/M fraction from noninoculated and fungus-inoculated roots revealed 36 membrane-associated protein modifications related to the functional AM symbiosis, including 15 induced, three up-regulated, and 18 down-regulated spots. Among the 25 mycorrhiza-responsive proteins that could be identified following MALDI-TOF and/or MS/MS analyses, 19 were not found to be regulated by a phosphate supply, suggesting they might be candidates as markers of AM symbiosis. Among them were an up-regulated 53-kDa nodulin and an up-regulated acid phosphatase along with a down-regulated lipoxxygenase fragment. This was the first report concerning the overexpression of a nodulin 53-like protein in mycorrhizal roots. Because this nodulin was found to be located in the peribacteroid membrane of root nodules [27], it was suggested that this protein could also be associated with the periarbuscular membrane. Regarding the acid phosphatase, the increase in the amount and activity of a similar enzyme also found to occur during nodule development was proposed to be important for efficient nodule metabolism [40]. The decreased abundance in a lipoxxygenase fragment suggested that this enzyme might be down-regulated during AM symbiosis. The lipoxxygenase pathway is induced during some plant–pathogen interactions leading to the production of signaling molecules, activation of defense genes, and phytoalexin accumulation [41]. Earlier studies have shown that defense-response genes involved in phytoalexin biosynthesis were down-regulated as the symbiosis developed [42].

Because very little is known about the mycorrhizal periarbuscular membrane, attempts to purify this compartment have relied on its plant-derived PM origin. Due to the under-representation of intrinsic plant membrane proteins on 2D gels, strategies have thus been developed in order to ensure a more thorough coverage of plant PM proteins. With regard to AM symbiosis, a qualitative comparative proteomic approach of the PM fractions of *M. truncatula* roots colonized or not with *G. intraradices* was set up [43]. After root microsome extractions, a discontinuous sucrose gradient-based protocol was used to enrich to a similar extent the PM fractions from control and

mycorrhizal roots. To gain access to an enlarged inventory of the proteins present in the PM fractions, two complementary proteomic methodologies were used: (i) liquid trypsin digestion of the PM fraction and separation of the generated peptides by 2D LC-MS/MS and (ii) systematic LC-MS/MS analyses of SDS-PAGE resolved proteins. The PM fractions of two independent control root experiments were analyzed by the two above-cited methods resulting in the identification of 78 proteins. Fifty-six percent of them were predicted to contain one or more TM domains. Additionally, among the 34 proteins with no TM domain, five contained a predicted C-terminal peptide signal for GPI-anchorage to the PM. To detect the proteins only present in the PM fraction from *G. intraradices*-inoculated roots, the complete list of nonredundant proteins identified from the two control analyses was used as the reference protein database, and only proteins identified with at least four peptides in the PM fraction of inoculated roots were taken into account. Using these criteria, two proteins, corresponding to a PM proton-efflux P-type ATPase (*Mth1*) and a *Bcp*, were found induced in mycorrhizal roots as compared to control ones. *Mth1* is a highly intrinsic protein containing 10 possible TM domains whose activity generates an electrochemical H^+ gradient across the cell PM using ATP. PM proton-efflux P-type ATPases are believed to produce the driving force necessary for the uptake and efflux of solutes across the plant-microbe interface. Because the expression of *Mth1* was previously found to be located in arbuscule-containing cells [44], it was suggested that *Mth1* was likely to be located in the periarbuscular membrane. H^+ -ATPases also have been detected in the symbiosome membrane of root nodules [45]. By contrast, ATPase activity is generally lacking from the symbiosome membranes of pathogenic interactions [3]. The *Bcp* is a small protein characterized by a copper-binding site together with a putative signal for GPI anchoring at the PM. GPI-anchored proteins also are known to localize in cholesterol- and glycosphingolipid-rich domains on the cell surface called lipid rafts that have been implicated in microbe-host cell interactions and infections [46]. A GPI-anchored nodulin (ENOD16) has been detected in the plant-derived symbiosome membrane of *M. truncatula* interacting with *S. meliloti* [45]. The *Bcp* encoding gene has been previously reported as induced in AM-inoculated roots by *in silico* EST analyses and cDNA arrays [47, 48]. By the use of reporter gene expression, it was also demonstrated that *Bcp* expression was specifically up-regulated in arbuscule-containing regions of mycorrhizal roots, suggesting its possible location in the periarbuscular membrane [49].

Proteomics at the Appressorium Stage

In contrast to many plant-pathogenic fungus interactions [50] very little is known about the molecular events occurring prior to and during the initial contact between the two symbionts. Modification of the *M. truncatula* root proteome during the early stage of AM symbiosis was recently investigated by comparing, using 2-DGE, the protein patterns obtained from noninoculated roots and roots synchronized for *G. intraradices* appressorium formation [51]. Statistically significant changes in protein abundance were recorded between inoculated and noninoculated roots, showing pre-infection-triggered root proteome modifications. The use of MALDI-TOF-MS

enabled the first identifications of plant root proteins whose accumulation was altered in response to early stages of AM symbiosis. Among the six root proteins that displayed an increase in abundance in *G. intraradices*-inoculated J5 wild-type plants, a chalcone reductase (CHR), a glutathione-dependent DHAR, a cyclophilin (CYP), and an actin depolymerization factor (ADF) were identified. Root proteins that decreased in density upon appressorium formation corresponded to two GST. The identification in the wild-type genotype of a CHR that increased in abundance prior to root penetration was consistent with the hypothesis that activation of the phenylpropanoid pathway is mediated by early recognition events between the two symbionts [52]. In view of the decreased abundance of two GSTs in response to *G. intraradices*, it was also suggested that some plant defense reactions might be lowered upon symbiont contact. Because a role in signal transduction has been proposed for the DHAR, CYP, and ADF that displayed an increased accumulation in *G. intraradices*-inoculated wild-type roots, this study also pointed out the elicitation of new signaling proteins in response to appressorium formation.

Major differences were also observed when comparing the proteins that responded to appressorium formation between the wild-type genotype and the *sunn* mutant defective for the autoregulation of arbuscule formation (hyper-mycorrhizal phenotype) [51]. They included changes in the abundance of a CHR, a narbonin, two annexins (MtAnn1), a quinone reductase, a protease inhibitor, and DHAR, CYP, and ADF proteins. This result indicated that differences in protein accumulation occurring prior to root penetration preceded the overdevelopment of fungal arbuscules in the hyper-mycorrhizal mutant. Because earlier studies have reported increased *MtAnn1* expression [53] and decreased levels of narbonin, Kunitz-type trypsin inhibitor, and quinone reductase transcripts in the roots of the *sunn* mutant upon *S. meliloti* inoculation [54], the results from this proteomic study also pointed out common early regulatory pathways that may be shared between the two symbioses. From the identification of proteins whose response to appressorium formation was modified in the *M. truncatula sunn* mutant, it was also suggested that defense reactions, cytoskeleton rearrangement and/or membrane trafficking events, and auxin signaling might be involved in the early control of mycorrhization.

42.4 A 2005–2006 UPDATE OF THE CONTRIBUTION OF PROTEOMICS TO RNF SYMBIOSIS

Contrasting with earlier studies that mainly concentrated on symbiosome-related proteins [22, 27, 45, 55], most recent insights in the proteomic analysis of RNF symbiosis come from experiments addressing early stages of infection. The dominant analytical device has been 2-DGE coupled to MS, but proteome modifications upon rhizobium inoculation or Nod factor application were investigated at different levels of complexity, including whole root tissues, isolated root hair cells, and root microsomes. With regard to root nodules, two additional studies, detailed below, also shed additional light on the protein analysis of bacteroids and their symbiosome membrane.

Proteomics of the Early Stages

In RNF symbiosis, the host plant also maintains a certain number of nodules on the root system by regulating locally and systemically the initiation of both infection threads and nodule primordia. In *M. truncatula*, an ethylene-insensitive mutant (*skl*) that displays a supernodulating phenotype when inoculated with *S. meliloti* has been isolated, and it has been suggested that the increase in nodule numbers and sustained infection thread growth in *skl* could be caused by the inhibition of some plant defense reactions [56].

To investigate the molecular details about the action of ethylene during early nodule development in *M. truncatula*, Prayitno et al. [57] used a comparative proteomics approach based on 2-DGE, image analysis, and MALDI-TOF-MS and/or LC-MS/MS. When comparing the proteome of *skl* roots to its wild-type, one and three days post inoculation, six proteins (pprg-2, Kunitz proteinase inhibitor, and ACC oxidase isoforms) were down-regulated in *skl* roots, while three protein spots were up-regulated (trypsin inhibitor, albumin 2, and CPRD49). To study ethylene-mediated protein changes in root tissues, the root proteome of the *skl* mutant to its wild-type in response to the ethylene precursor ACC also was compared. The authors demonstrated that ACC induced stress-related proteins in wild-type roots, such as pprg-2 protein, ACC oxidase, proteinase inhibitor, APX, and HSPs. However, the expression of pprg-2, Kunitz proteinase inhibitor, and ACC oxidase was down-regulated in inoculated *skl* roots, indicating that these stress-related proteins were regulated by ethylene at early stages of nodule formation. It has thus been hypothesized that during early nodule development, the plant induces ethylene-mediated stress responses to limit infection and nodule numbers. When a mutant defective in ethylene signaling, such as *skl*, is inoculated with rhizobia, the plant-stress response is reduced, resulting in increased nodule numbers.

The proteome of soybean root hairs after infection with *B. japonicum* also has been recently investigated [25]. Root hairs are the extensions of root epidermal cells in the elongation zone and are the primary sites for rhizobial infection. Because they are morphologically and also probably biochemically different from the remaining epidermal cells and roots, their response to rhizobial infection also may be different from that of roots in terms of protein species, induction level, and kinetics. To investigate this point, 2-DGE total protein profiles obtained from noninoculated root and root hairs were first compared. Among the proteins that were unique or more abundant in root hairs were a CaM-like domain protein kinase showing similarity to DMI3, a protein required for the formation of both RNF and AM symbioses, and a phosphoenolpyruvate carboxylase that can produce dicarboxylic acids, a preferred carbon source for rhizobial bacteroids [3]. 2-DGE protein profiles of root hairs treated with *B. japonicum* or water as a control were also compared to identify root-hair proteins responsive to the symbiont. Among the 26 spots that appeared to be differentially expressed between the two treatments, 16 spots could be identified following MS/MS analyses. They included proteins previously known to respond to rhizobial inoculation (lipoxygenase, PAL, and L-APX), but also additional ones that matched to PLD, phosphoglucosucrose, and actin proteins. PLD might play an important role in symbiosis initiation through its function in lipid signaling pathway and microtubule

reorganization. Likewise, the induction of actin expression upon inoculation fitted with the cytoskeleton rearrangements that are known to occur at the root-hair tip during the early stages of infection. Regarding phosphoglucosylases, it has been proposed that they may affect the availability of host phosphoglucose to the rhizobium.

Because the PM acts as a communication interface with the extracellular environment for the exchange of information and metabolites, Boukli et al. [58] investigated PM protein changes in response to Nod-factor application to roots of *V. unguiculata*. For that purpose, PMs were purified from the root microsomal fraction by using a dextran-polyethylene glycol two-phase partition protocol. Comparison of PM 2D gel protein patterns revealed the induction of a 57-kDa protein in Nod-factor treated roots. Following N-terminal sequencing, this protein was identified as a V-ATPase that showed high sequence similarity to that of *A. thaliana*. V-ATPases are a family of ATP-dependent ion pumps that play central role in endocytic and secretory trafficking. In plants, they have been localized to vacuoles and other membranes of the secretory system, including the ER, PM, Golgi, small vesicles, and the peribacteroid membrane of nodules [27]. It has been proposed that the induction of a V-ATPase in roots upon Nod factor application may be associated with cell swelling and elongation [58].

Proteomics of Root Nodules

Transformation of a free-living bacterium to a nitrogen-fixing bacteroid results in significant physiological and developmental changes in rhizobia. To understand the cellular events occurring in bacteroids isolated from nitrogen-fixing root nodules, a 2D gel proteome map was constructed using the total protein extracts from *B. japonicum* bacteroids [59]. Among the 400 protein spots that could be detected, 180 were identified using MALDI–TOF–MS by searching the available database for *B. japonicum*. The data showed that the bacteroid expressed a dominant and elaborate network for nitrogen and carbon metabolism, which is closely dependent on the plant-supplied metabolites. Although the *B. japonicum* genome encodes over 700 genes related to transport, only a number of specific ABC transporters appeared to be expressed in bacteroids. They also seemed to lack a defined fatty acid and nucleic acid metabolism. Proteins related to protein synthesis, scaffolding, and degradation were among the most abundant spots of the bacteroid proteome. Additionally, several proteins were identified as involved in cellular detoxification and stress regulation. Among them was a single prominent SOD. This enzyme that scavenges ROS has been shown to be important in the symbiotic process [60]. Interestingly, the presence in AM fungus of a functional SOD gene that is modulated during the fungal life cycle also has been proposed to offer protection as a ROS-inactivating system against localized host defense responses raised in arbuscule-containing cells [61]. Likewise, a putative fungal APX that may be involved in combating a plant-induced oxidative burst also was recently detected in the proteome of wheat leaves upon infection with the pathogenic biotrophic fungus *P. tritici* [15]. Overall, this proteomic analysis of *B. japonicum* bacteroids pointed out the interconnection between several metabolic pathways.

The second study that recently addressed root nodules exemplifies very elegantly the further in-depth analysis of candidate proteins identified through proteomics.

When investigating the symbiosome membrane proteins in the *M. truncatula*-*S. meliloti* interaction using 2-DGE and LC-MS/MS, Catalano et al. [45] identified, along with many proteins related to protein destination and storage, a putative *M. truncatula* syntaxin (MtSYP132). MtSYP132 is highly homologous to one of the nine-member SYP1 groups of *Arabidopsis* syntaxins of unknown function. Syntaxins belong to the SNARE superfamily of proteins that are central to the process of secretory vesicle targeting and fusion. Additionally, SNARE proteins are known to be enriched in detergent resistant membranes [46]. To evaluate the distribution of MtSYP132 in root nodule tissue, Catalano et al. [62] used a specific antibody generated against a peptide antigen from the predicted sequence of MtSYP132. Total nodule, total nodule membrane, symbiosome membrane and space, and bacteroid protein fractions were further probed with anti-MtSYP132. Cytological analyses along with Western blotting experiments demonstrated that MtSYP132 was localized specifically on the PM surrounding the infection thread and droplet and not on other PM within the same cell. It was suggested that protein changes in the PM surrounding the infection thread are important for defining the PM region that becomes a functional infection droplet. Interestingly, among the plant-unique roles attributed to syntaxins is their involvement in mediating some plant pathogen responses, including disease resistance at the plant cell wall [46].

42.5 CONCLUSIONS

During the last two years, the use of proteomics approaches more specifically directed toward well-defined stages/tissues of root endosymbioses and/or subcellular fractions has enabled the identification of additional SMRPs. Interestingly, many of the above-listed proteins tend to support the existence of overlaps between the developmental programs of AM and RNF symbioses (Table 42.1). The proteins identified when analyses were focused on early stages of symbioses were consistent with the transient activation of the phenylpropanoid pathway upon inoculation with AM fungi (CHR) and rhizobia (PAL). Major insights came from the proteomic analyses of autoregulation-defective mutants relative to their wild-type counterparts, which pointed out the involvement of early plant defense reactions to limit nodule and arbuscule numbers. Overlapping plant responses to AM fungi and rhizobia in the *sunn* mutant also encompassed proteins related to cytoskeleton rearrangements and auxin signaling. When regarding proteomics studies devoted to symbiosomes, overlapping patterns of accumulation between AM and RNF symbioses were observed for a 53-kDa nodulin and an acid phosphatase of plant origin but of unknown function. Likewise, GPI-anchorage in putative lipid rafts of the PM has been detected for the proteins Bcp and ENOD16 in AM and RNF symbioses, respectively. Overall, it seems that intracellular symbionts modulate the protein composition of plant endomembranes.

42.6 FIVE-YEAR VIEWPOINT

Despite the importance of protein phosphorylation in signal transduction during cell defense responses in PMIs, phosphoproteomics approaches so far only have been

TABLE 42.1. Examples of Additional Putative Proteins Identified Using Proteomics that Show Overlapping Patterns of Accumulation Between AM and RNF Symbioses

Plant Proteins	AM Symbiosis	RNF Symbiosis
Early Symbiosis		
Wild-Type Genotype:		
Phenylpropanoid pathway	Amiour et al. [51]	Wan et al. [25]
Actin-related proteins	Amiour et al. [51]	Wan et al. [25]
<i>sun1</i> Mutant:		
MtAnn1	Amiour et al. [51]	Carvalho-Niebel et al. [53]
Kunitz-type trypsin inhibitor	Amiour et al. [51]	El Yahyaoui et al. [54]
Quinone reductase	Amiour et al. [51]	El Yahyaoui et al. [54]
Narbonin	Amiour et al. [51]	El Yahyaoui et al. [54]
Functional symbiosis		
Wild-Type Genotype:		
53-kDa nodulin	Valot et al. [39]	Salbaach et al. [27]
Acid phosphatase	Valot et al. [39]	Peinheiter et al. [40]
H ⁺ -ATPase	Valot et al. [43]	Catalano et al. (2004)
GPI-anchored proteins	Valot et al. [43]	Catalano et al. (2004)
Microbial Proteins:		
SOD	Lanfranco et al. [61]	Sarma and Emerich [59]

developed for pathogenic associations [12]. The progress made in the analysis of phosphopeptides should make possible the identification of new components of the regulatory pathways involved in the establishment of AM and RNF symbioses. Likewise, changes in membrane and cell wall proteins also are likely to occur when microsymbionts contact their hosts, so that the development of subcellular proteomics targeting the early stages of symbioses is expected to be very fruitful. When regarding proteomic studies devoted to symbiosomes, recent technical developments in MS analyses are expected to provide new insights about the intrinsic proteins that are regulated in AM and RNF symbioses. Methods such as ICAT labeling that permit to compare peptide abundance between two samples appear well-suited to enable a quantitative comparison of membrane proteins between symbiotic and nonsymbiotic roots. The possibility to detect GPI-anchored proteins by their enzymatic release from PM fractions using phospholipases also will allow their high-scale characterization during symbiosis functioning [13]. Although it is still unknown whether it exists as core protein(s) in the plant-derived membrane that surrounds intracellular biotrophic microorganisms, *M. truncatula* appears well-suited to engage comparative proteomic studies targeting perihyphal, periarbuscular, and peribacteroid membranes (<http://www.noble.org/MedicagoHandbook>). The use of LCM-based techniques that can be coupled to proteome profiling also will help to enrich samples in fungal structures. Regarding proteins of plant origin, transformation methods coupled to protein labeling and microscopy are among the strategies that will help to validate their putative location within membranes. The availability of plant mutants generated

through Targeting Induced Local Lesions in Genomes (TILLING) and/or insertional mutagenesis will be very useful for the functional testing of biotrophy-related candidate proteins.

ACKNOWLEDGMENT

The authors thank the Council of Burgundy for financial support (04 516 CP09 5327).

REFERENCES

1. Garcia-Brugger A, et al. 2006. *Mol Plant-Microbe Interact* **19**:711–724.
2. Pazkowski U. 2006. *Curr Opin Plant Biol* **9**:364–370.
3. Parniske M. 2000. *Curr Opin Plant Biol* **3**:320–328.
4. O'Connell RJ and Panstruga R. 2006. *New Phytol* **171**:699–718.
5. Schulze-Lefert P and Panstruga R. 2003. *Annu Rev Phytopathol* **41**:641–667.
6. Smith SE and Smith A. 1990. *New Phytol* **114**:1–38.
7. Harrison MJ. 1999. *J Exp Biol* **50**:1013–1022.
8. Perfect SE and Green Jr. 2001. *Mol Plant Pathol* **2**:101–108.
9. Mathesius U. 2001. *Plant Soil* **255**:105–119.
10. Panstruga R. 2003. *Curr Opin Plant Biol* **6**:320–326.
11. Canovas FM, et al. 2004. *Proteomics* **4**:285–298.
12. Thurston G, et al. 2005. *Physiol Mol Plant Pathol* **66**:3–11.
13. Rossignol M, et al. 2006. *Proteomics* **6**:5529–5548.
14. Curto M, et al. 2006. *Proteomics* **6**:S163–S174.
15. Rampitsch C, et al. 2006. *Proteomics* **6**:1897–1907.
16. Bestel-Corre G, Dumas-Gaudot E, and Gianinazzi S. 2004. *Mycorrhiza* **14**:1–10.
17. Schüßler A, Schwarzott D, and Walker C. 2001. *Mycol Res* **105**:1413–1421.
18. Bonfante P and Perotto S. 1995. *New Phytol* **130**:3–21.
19. Remy W, et al. 1994. *Proc Natl Acad Sci USA* **91**:11841–11843.
20. Doyle JJ. 1998. *Trends Plant Sci* **3**:473–478.
21. Oldroyd G, Harrison M, and Udvardi M. 2005. *Plant Physiol* **137**:1205–1210.
22. Djordjevic MA. 2004. *Proteomics* **4**:1859–1872.
23. Verma D and Hong Z. 1996. *Trends Microbiol* **4**:364–368.
24. Balestrini R and Lanfranco L. 2006. *Mycorrhiza* **16**:509–524.
25. Wan J, et al. 2005. *Mol Plant-Microbe Interact* **18**:458–467.
26. Weidmann S, et al. 2004. *Mol Plant-Microbe Interact* **17**:1385–1393.
27. Saalbach G, Erik P, and Wienkoop S. 2002. *Proteomics* **2**:325–337.
28. He F, et al. 2004. *J Proteome Res* **3**:61–67.
29. Liska AJ, et al. 2004. *Proteomics* **4**:S2707–S2721.
30. Cook DR. 1999. *Curr Opin Plant Biol* **2**:301–304.
31. Stacey G, et al. 2006. *Curr Opin Plant Biol* **9**:110–121.

32. Güimil S, et al. 2005. *Proc Natl Acad Sci USA* **102**:8066–8070.
33. Martin F, et al. 2004. *New Phytol* **161**:330–335.
34. Puhler A, et al. 2004. *Curr Opin Plant Biol* **7**:137–147.
35. Kaneko T, et al. 2000. *DNA Res* **7**:331–338.
36. Dumas-Gaudot E, et al. 2004. *FEMS Microbiol Ecol* **48**:401–411.
37. Bestel-Corre G, et al. 2002. *Electrophoresis* **23**:122–137.
38. Valot V, Dumas-Gaudot E, and Gianinazzi S. 2004. *Phytochemistry* **65**:1721–1732.
39. Valot B, et al. 2005. *Plant Mol Biol* **59**:565–580.
40. Penheiter AR, Duff SM, and Sarath G. 1997. *Plant Physiol* **114**:597–604.
41. Blee E. 2002. *Trends Plant Sci* **7**:315–321.
42. Harrison M and Dixon RA 1993. *Mol Plant–Microbe Interact* **6**:643–654.
43. Valot B, et al. 2006. *Proteomics* **6**:S145–S155.
44. Krajinski F, et al. 2002. *Plant Biology* **4**:754–761.
45. Catalano CM, Lane WS, and Sherrier DJ. 2004. *Electrophoresis* **25**:519–531.
46. Bhat RA and Panstruga R. 2005. *Planta* **223**:5–19.
47. Journet EP, et al. 2002. *Nucleic Acids Res* **15**:5579–5592.
48. Manthey K, et al. 2004. *Mol Plant–Microbe Interact* **17**:1063–1077.
49. Hohnjec N, et al. 2005. *Plant Physiol* **137**:1283–1301.
50. Tucker SL and Talbot NJ. 2001. *Annu Rev Phytopathol* **39**:385–417.
51. Amiour N, et al. 2006. *Mol Plant–Microbe Interact* **19**:988–997.
52. Garcia-Garrido JM and Ocampo JA. 2002. *J Exp Biol* **53**:1377–1386.
53. Carvalho-Niebel F, et al. 1998. *Mol Plant–Microbe Interact* **11**:504–513.
54. El Yahyaoui F, et al. 2004. *Plant Physiol* **136**:3159–3176.
55. Natera SH, Guerreiro N, and Djordjevic MA. 2000. *Mol Plant–Microbe Interact* **13**:995–1009.
56. Penmetsa R and Cook D. 2000. *Plant Physiol* **123**:1387–1398.
57. Prayitno J, et al. 2006. *J Proteome Res* **5**:3084–3095.
58. Boukli NM, et al. 2007. *J Plant Physiol* **164**:794–806.
59. Sarma AD and Emerich DW. 2005. *Proteomics* **5**:4170–4184.
60. Santos R, et al. 2000. *Mol Microbiol* **4**:750–759.
61. Lanfranco L, Novero M, and Bonfante P. 2005. *Plant Physiol* **137**:1319–1330.
62. Catalano CM, et al. 2007. *Planta* **225**:541–550.

PROTEOMICS APPROACHES TO CONSTRUCT CALCIUM SIGNALING NETWORKS IN PLANTS

Irene S. Day and A. S. N. Reddy

43.1 INTRODUCTION AND BRIEF BIBLIOGRAPHIC REVIEW

Plant growth and development is controlled by many internal signals such as hormones and external signals including many biotic and abiotic signals. Plants have developed mechanisms to perceive these signals and to produce an appropriate response. Receptors perceive the signals and transmit this information to cellular machinery through non-protein messengers, enzymes, and/or transcription factors. The nonprotein messengers include organic molecules such as cyclic nucleotides and hydrolysis products of phosphoinositides, as well as inorganic molecules/ions such as H_2O_2 , NO, and Ca^{2+} [1, 2]. The cytoplasmic concentration of calcium ion [Ca^{2+} ($[\text{Ca}^{2+}]_{\text{cyt}}$)] increases in response to many developmental and hormonal cues and biotic and abiotic stimuli [3]. Calcium is thought to be widely used in signal transduction. Calcium-mediated signal transduction has been widely studied and has been implicated in diverse processes including pollen tube growth, root-hair and lateral root development, nodulation, cell division and elongation, stomatal closure/opening, pathogen- and elicitor-induced defense, and abiotic stress responses [2]. Calcium signaling involves numerous proteins such as channels, pumps, transporters, sensors, and targets. Characterizing these

components on a one-protein-at-a-time basis is slow and laborious. High-throughput proteomics approaches for studying calcium-signaling networks will greatly facilitate the identification of the various components and, more importantly, interactions between the components in a more efficient manner.

Components of Calcium Signaling

Calcium signaling has three major phases (generation of Ca^{2+} signature, recognition of Ca^{2+} changes and generation of an appropriate response) (Figure 43.1). Channels, pumps, and transporters are involved in generating a change in $[\text{Ca}^{2+}]_{\text{cyt}}$ specific to the signal that is termed the Ca^{2+} signature [4]. The change in $[\text{Ca}^{2+}]_{\text{cyt}}$ is then sensed by calcium-binding proteins (calcium sensors), and these proteins in turn regulate the activity/function of target proteins involved in producing a response to the signal (Figure 43.1).

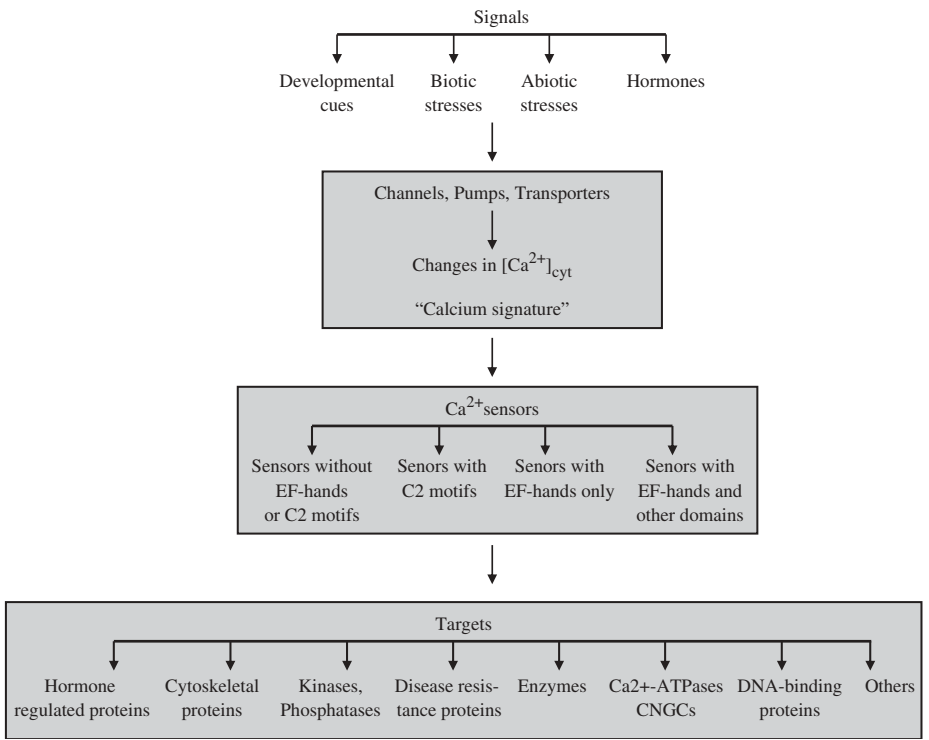


FIGURE 43.1. The three phases in calcium signaling. The first phase is the generation of the calcium signature, the second phase is sensing of the change in Ca^{2+} by calcium sensors, and the third phase is interaction of the sensors with targets, resulting in the response to the signal. CNGC, cyclic nucleotide gated channels.

Generation of the Calcium Signature

The $[Ca^{2+}]_{cyt}$ in resting cells is 100–200 nM, while in extracellular and intracellular Ca^{2+} stores it is thought to be in the millimolar range (1–10 mM). The $[Ca^{2+}]_{cyt}$ levels are regulated in response to signals by channels, pumps, and transporters and can be elevated up to 3 μ M, with the magnitude and the duration varying depending on the signal or cell type. Fifty-one potential channels and 30 potential pumps and transporters have been identified in *Arabidopsis*, suggesting complexity in the regulation of signal-specific calcium signatures [2]. However, the role of these proteins in regulating $[Ca^{2+}]_{cyt}$ is poorly understood. Although there are a few published reports on individual proteins involved in production of the calcium-signature, much more work needs to be done to unravel the role of these channels, pumps, and transporters and their specificities that result in $[Ca^{2+}]_{cyt}$ levels that are varied in response to different stimuli.

Calcium Sensors

Changes in the calcium concentration in the cytosol and subcellular compartments are sensed by calcium-binding proteins. In most cases, Ca^{2+} binds to a conserved motif in the calcium sensors. The most common motif is an EF-hand motif consisting of a 29-amino-acid helix–loop–helix. The helices termed E and F flank the calcium-binding loop (Figure 43.2A). Some EF-hand proteins—that is, CaMs and CaM-like (CML) proteins—have no other recognizable domains [5]. Other EF-hand proteins contain various domains that indicate their involvement in several cellular processes including signal relays, phosphorylation and dephosphorylation cascades, production of second messengers, ion transport pathways, transcription and translation, phospholipids-based signaling, and membrane trafficking [2]. Two other motifs found in plant proteins that mediate Ca^{2+} -dependent interaction of proteins with membranes or membrane lipids are an annexin fold in members of the membrane-associated annexin subfamily and the C2 domain, a domain of about 130–145 amino acids present in other membrane-associated proteins. In addition to these, some proteins that were shown experimentally to bind Ca^{2+} do not have any of the three conserved motifs, suggesting the existence of other novel types of calcium sensors.

Calcium Sensor Targets

In most cases, binding of Ca^{2+} to these sensors alters their confirmation, resulting in their interaction with other proteins (Ca^{2+} sensor target proteins) and regulation of these target proteins' activities. The large variation in calcium-binding motifs, the number of motifs, and the presence of other domains in calcium sensors suggest a large number of target proteins that interact with the calcium sensors. Calcium sensors with no other domain are thought to activate/inactivate their target proteins, while other sensors such as CDPKs can act directly on their targets. Some targets have been identified using PPI-based approaches, Y2H assays or *in vitro* protein kinase assays [2].

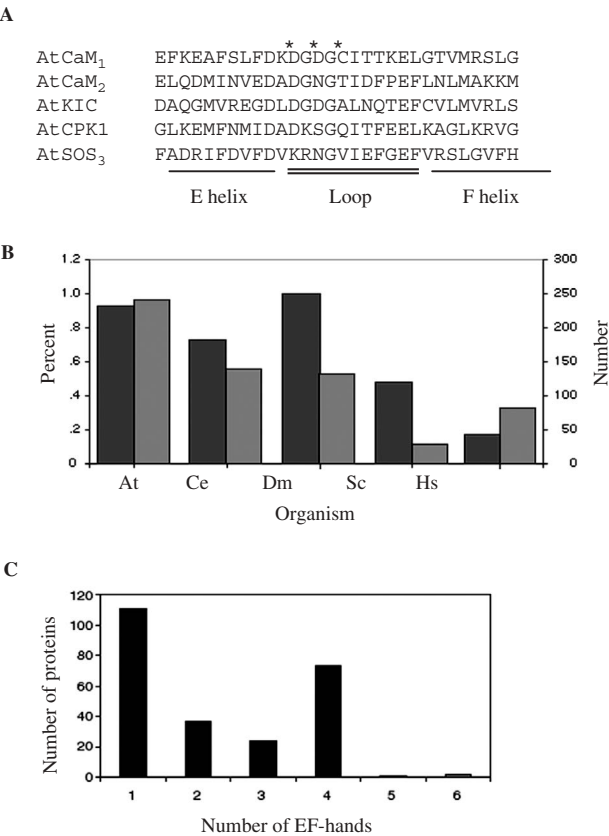


FIGURE 43.2. EF-hand proteins in *Arabidopsis*. **(A)** An alignment of the EF-hand motif from the first and second EF-hands in *Arabidopsis* CaM₁, the one EF-hand motif in AtKIC, and the first EF-hand in AtCPK1 (a CDPK) and AtSOS3 (a CBL). Asterisks (*) denote amino acids conserved in all five EF-hands. **(B)** Graph showing the percent (*left*) of the proteome represented by and the number (*right*) of EF-hand proteins in *Arabidopsis* (At), *C. elegans* (Ce), *Drosophila* (Dm), *S. cerevisiae* (Sc), and humans (Hs). **(C)** The number of proteins in *Arabidopsis* containing the indicated number of EF-hand motifs. B and C are adapted from Day et al. [6].

Protein Components in Calcium Signaling Networks

All three phases in calcium signaling involve a large number of protein components. The number of proteins increases from phase 1 to 3, with the number of proteins involved in signal generation being smaller than the much larger number of known and predicted sensor proteins. The number of target proteins is expected to be much larger than the sensors, because each sensor can potentially interact with multiple proteins. This results in a vast number of proteins, protein interactions, and processes that need to be investigated in order to unravel calcium-signaling networks. Proteomics approaches offer high-throughput methods to achieve this goal.

43.2 METHODOLOGY AND EXPERIMENTAL RESULTS

Identification of Calcium Sensors

Bioinformatics Approaches. Sequencing of the complete genomes of many eukaryotic organisms and the availability of searchable databases has led to bioinformatic approaches for the identification of protein families. Highly conserved motifs in the families can be used to search the databases for potential family members in individual species and between species. At SMART (<http://smart.embl-heidelberg.de>) a taxon can be searched for a particular domain. A list of proteins containing that domain in the selected group can be retrieved. However, possibly not all returned proteins have the domain of interest and not all proteins that have the domain are present in the database. Therefore, using the list as a starting point, all proteins can be analyzed at InterProScan (www.ebi.ac.uk/InterProScan/) to verify the presence of the domain and also to identify other domains in each protein. To ensure that all the proteins in the genome of interest have been identified, it is necessary to do BLAST searches at NCBI (www.ncbi.nlm.nih.gov) or at websites for the particular genome of interest. Proteins identified in this search can also be verified and analyzed for other domains using InterProScan. The other domains in the identified proteins provide valuable clues as to the possible function of these proteins and can help guide the follow-up research.

Using domain prediction programs and BLAST searches with known calcium-binding proteins containing EF-hands [a CaM4, a calcineurin B-like protein (CLB3), a CPK1 and KCBP interacting calcium-binding protein], 250 proteins were identified in *Arabidopsis* predicted ORFs. These proteins contain one or more (up to six) EF-hand motifs [6] (Figure 43.2B, C). In *Arabidopsis* there are seven CaMs and 50 CML proteins, 10 CBLs, 34 CPKS, and many novel EF-hand containing proteins adding to the complexity of the calcium signaling network [5–7]. Each sequence was further analyzed using InterProScan to identify domains other than the EF-hand domain and to verify the presence and number of EF-hands in each protein. Some EF-hand-containing proteins have domains that indicate that they are involved in binding other proteins or other molecules including DNA, some have enzymatic domains, some domains indicate that the protein is involved in transcription or translation, and still others are identified as carriers or channels. Researchers interested in these varied processes can continue to study the function and regulation of these identified proteins. The C2 calcium-binding motif is also highly conserved. SMART identifies 122 C2 proteins in the *Arabidopsis* genome. These have yet to be analyzed for other domains and for the ability to bind calcium. The proteins that have been shown experimentally to bind Ca^{2+} but that do not have a known calcium-binding motif do not lend themselves to the same type of database searches.

Comparative analyses of calcium-binding proteins across species can be made using the databases from other sequenced genomes. Using the known calcium-binding proteins from one species to search for similar proteins in another species can identify putative calcium-binding proteins in the new group. Nagata et al. [8] searched the *Arabidopsis* and rice (*indica*) genomic sequence data and 32K full-length cDNAs from rice (*japonica*) with known animal calcium-binding proteins. They concluded

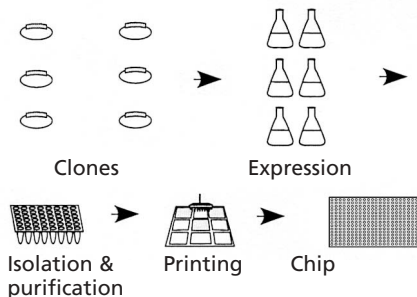
that voltage-dependent calcium ion channels were markedly different between plants and animals but other transport proteins were conserved. Many animal calcium-binding genes were not present in plants; conversely, there were many plant-specific genes.

High-Throughput Experimental Approaches. In recent years, DNA microarrays have become a powerful tool for high-throughput analysis of genomes and gene expression. By searching available expression databases, expression patterns of genes identified in the bioinformatics searches can be elucidated. McCormack et al. [5] used data obtained by massively parallel signature sequencing and compiled from *Arabidopsis* Affymetrix microarray chip experiments at Genevestigator (<https://www.genevestigator.ethz.ch>) to analyze the expression patterns of the seven *Arabidopsis* CaMs and 50 CML genes. The seven CaMs were found to be highly and relatively uniformly expressed, while the CMLs showed differential expression over developmental stages, in various organs and in response to many different stimuli.

DNA microarrays are helpful for expression data. However, for identification of novel calcium-binding proteins or for verification of actual calcium-binding of putative calcium-binding proteins, the more recently developed protein arrays are more useful (see Box 43.1). The development of protein microarrays was limited by the

Box 43.1

Protein Chip Technology



Protein chips are slides that have sets of different proteins spotted on them. Proteins are expressed, isolated and purified and then spotted onto a slide containing an entire set of proteins. The expressed proteins commonly have a tag (GST, HisX6, etc.) for purification that can then be used for quantification of the protein printed on the slide. AMAs are prepared from antibodies prepared against different proteins or epitopes. The arrays can then be probed with complex mixtures of proteins to identify antigens and the binding relationships and conditions required for binding.

Functional arrays, on the other hand, are slides spotted with sets of expressed, isolated and purified functional proteins that have been spotted to a slide. The sets could be specific to an organism, a tissue, a function or other specified condition. The sets of proteins on the slides can be characterized using a wide range of biochemical activities and PPIs.

relatively more difficult process of protein expression and isolation. As the availability of large open-reading-frame clone libraries has increased and high-throughput protein-production methods have improved, many types of protein chips have become available for non-plant systems [9]. Protein microarrays have been used for studying PPIs, PTMs, small molecule binding, and enzyme–substrate reactions [10].

Although bioinformatics analysis led to identification of over 250 putative and known calcium-binding proteins in *Arabidopsis*, the calcium-binding activity for most of these proteins has not been shown experimentally. A protein chip assay should lend itself to high-throughput analysis of calcium binding. In order to evaluate the feasibility of using protein arrays for identifying calcium-binding proteins, a pilot project can be used. Clones for a subset of the identified calcium-binding proteins can be obtained from ABRC (<http://www.biosci.ohio-state.edu/~plantbio/Facilities/abrc>) and transferred to an expression vector (Figure 43.3A). The proteins are expressed, electrophoresed, and blotted to nitrocellulose membranes. Membranes are incubated in calcium overlay buffer (60 mM KCl, 5 mM MgCl₂, 10 mM imidazole HCl, pH 6.8) and then incubated in the same buffer with 1 μ Ci/L ⁴⁵Ca at room temperature. The

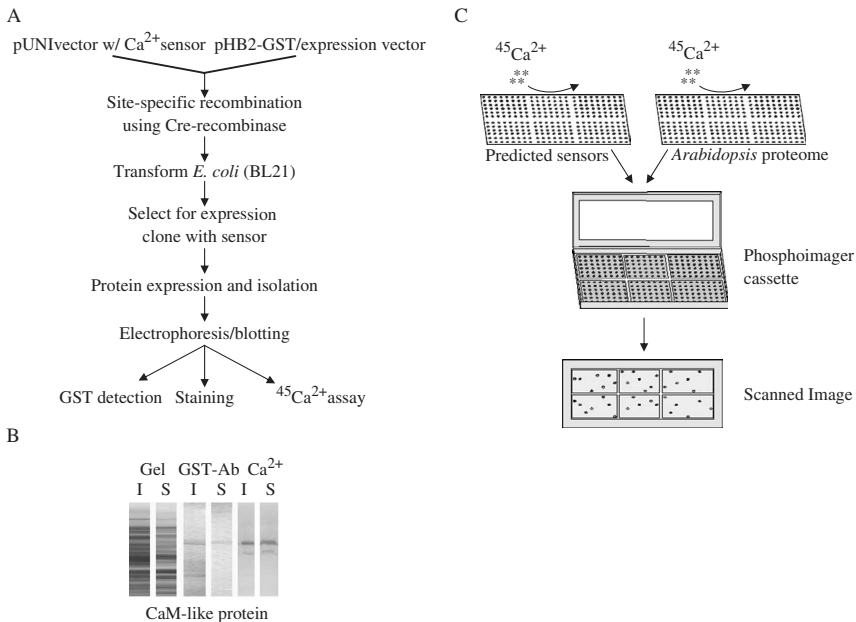


FIGURE 43.3. Ca²⁺ -binding assays. **(A)** Flow chart for cloning, protein expression, detection, and Ca²⁺ binding determination. **(B)** An example of the results for one Ca²⁺ sensor, a novel CaM-like protein. I, insoluble fraction; S, soluble fraction; Gel, stained gel; GST-antibody, detection of the fusion protein using GST antibody; Ca²⁺, calcium binding assay. **(C)** Proposed protein chip assay for identifying Ca²⁺ sensors. Chips containing either the predicted sensors or the *Arabidopsis* proteome are incubated in labeled Ca²⁺-containing buffer and the chip is then exposed to a phosphorimager screen, which is scanned to reveal the Ca²⁺-binding proteins. All dots on the scanned image represent hypothetical positive sensors.

membranes are washed in distilled water and 50% alcohol, exposed to a phosphor-imager screen for 4 h to overnight and scanned to identify proteins binding calcium. Using this approach, many of the bacterially expressed proteins we tested were positive for calcium-binding under the conditions used (Figure 43.3B). This experiment demonstrates the feasibility of using bacterially produced protein to investigate Ca^{2+} -binding. Two types of protein microarrays could be used to identify calcium-binding proteins. To verify the calcium binding of all the identified putative calcium-binding proteins, a chip could be constructed with the specific proteins identified in the database search and then probed with $^{45}\text{Ca}^{2+}$ (Figure 43.3C). The binding conditions, especially Ca^{2+} concentration, could be varied as calcium-binding conditions differ for different calcium sensors. The second type of assay would involve complete proteome chips. Probing these chips would identify calcium sensors that do not have a conserved calcium-binding domain as well as those whose calcium binding has been predicted. In this case a protein chip with proteins representing the proteome of an organism could be probed with the $^{45}\text{Ca}^{2+}$ under varied conditions (Figure 43.3C). These chip assays would not only identify calcium sensors but would also differentiate sensor affinity to Ca^{2+} .

Identification of Targets of Calcium Sensors

The diverse range of calcium sensors suggests that they are likely to interact with a large group of targets. Several approaches are available for identification of target proteins. Some of these have been used successfully to identify targets of some calcium sensors, especially CaM and CBLs.

Screening of Expression Libraries. Radioactive and nonradioactive methods for isolation of cDNAs encoding CBPs have been developed and successfully used for screening many expression libraries [11, 12]. Expression libraries can be plated and screened with either biotinylated CaM or ^{35}S -CaM (see Figure 43.4A for flow sheet). Positive colonies can then be plaque-purified using two more rounds of screening (Figure 43.4B) and the plasmid DNA is sequenced. Using this method, nearly 150 CBPs have been identified [12–15].

Y2H Screens. Y2H screens are another method for identifying calcium sensor targets by PPI. Individual Y2H assays have identified many target proteins. The Y2H system was used to identify CBL-interacting proteins and their interactions with the 18 kinases that are their targets [16]. However, screens with one clone/protein at a time are time-consuming and laborious. Global Y2H assays have been attempted using proteins from organisms ranging from yeast to humans [17, 18]. Complete proteomes have been used to construct all ORFs in bait and prey constructs to be used in genome-wide large-scale binary screens conducted using automated, robotic platforms. These screens have generated a large body of interaction data. Using binary interactive Y2H assays for a subset of the proteome such as the calcium sensors and their targets could identify interaction networks among these proteins. High-throughput Y2H assays with the calcium sensors as bait against all *Arabidopsis* ORFs as prey should

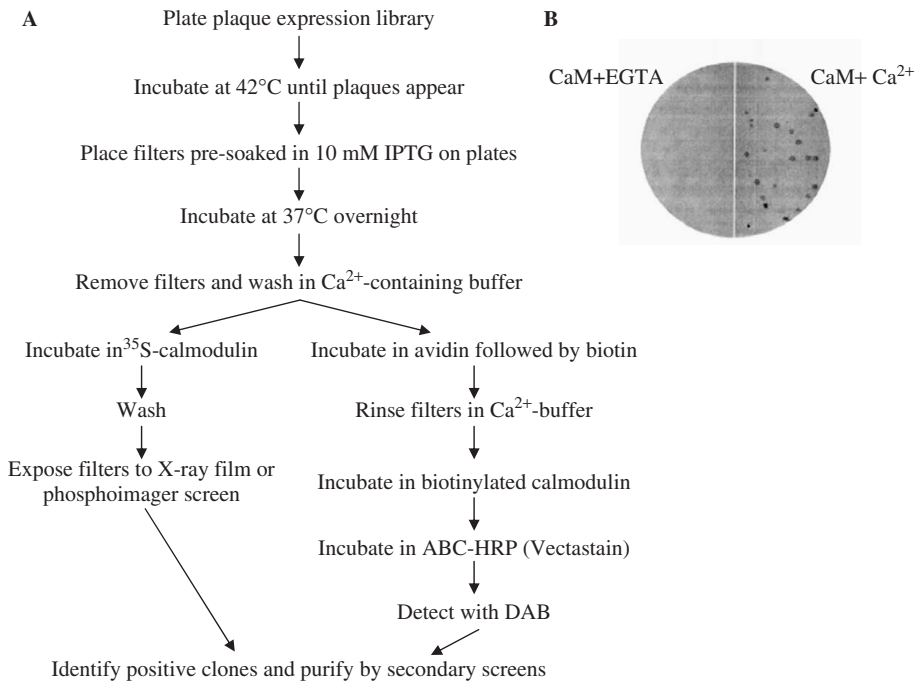


FIGURE 43.4. Screening of expression libraries for calcium sensor targets. **(A)** Flow sheet for screening an expression library with a labeled sensor. **(B)** An example of a filter from a screen showing binding of labeled CaM to a target protein in the expression library.

identify many sensor targets. However, the Ca^{2+} requirement of the bait proteins needs to be met in some way for these assays. This problem was overcome for a Y2H assay using a CDPK (AtCPK11) as bait [19]. The bait protein in this case was expressed as either constitutively active or catalytically inactive overriding the requirement for Ca^{2+} . Novel approaches like this need to be developed for effective use of calcium sensors as bait.

Using Protein Chips to Identify Calcium Sensor Targets and Interaction Networks between Calcium Sensors and Targets. Proteome chips are potentially the most useful method for identification of calcium sensor targets. The protein chips used for screening for calcium binding can be used for PPI screens and activity screens for enzymes (Figure 43.5B, C). Protein chips have been used for the global analysis of protein phosphorylation in yeast [20] and protein modification that involved covalent attachment of a small Ub-like modifier protein to target proteins [21]. A protein array containing 1690 proteins from *Arabidopsis* was used for identifying phosphorylation candidates for the protein kinases MPK3 and MPK6 [22]. From this small portion of the proteome (6.7%), 48 and 39 candidates respectively were identified. Because as there are 34 CDPKs in *Arabidopsis*, studying phosphorylation of

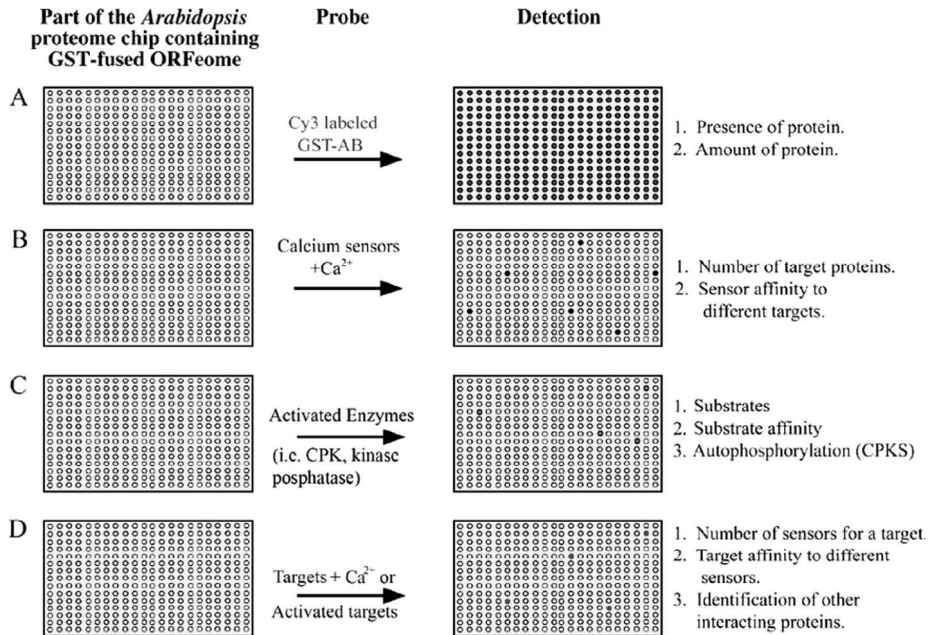


FIGURE 43.5. Proteome chip assays. Chips on the left contain proteins expressed from each ORF from *Arabidopsis*. Possible probes are given in the middle lane and mock detection chips on the right show some positive spots. The possible information gained is listed in the last column. **(A)** Probing with the GST antibody is for quality control, showing that the protein is on the chip and that there are equal amounts of each protein. **(B)** Known calcium sensors can be used as probes with varying concentrations of Ca²⁺ to identify targets and affinity to the targets at each Ca²⁺ concentration. **(C)** Sensors that have possible enzymatic functions can be used in enzymatic assays by providing the necessary conditions—that is, labeled ATP for CPK and kinase assays. Phosphorylated targets on the chip could be identified. For phosphatase assays, the proteins on the chip would need to be phosphorylated first and then incubated with the phosphatase. **(D)** Known sensor targets could be used as a probe with varying concentrations of Ca²⁺ to identify other sensors that might interact with the target and the affinity of the interaction. Or an activated target (in the presence of the interacting sensor and Ca²⁺) could be used as a probe to identify proteins (other than the sensor) that interact with it.

target proteins using protein chips would provide valuable information about autophosphorylation and the targets of the individual CDPKs. Phosphorylation assays could be performed with each CDPK using *Arabidopsis* proteomic chips, which should be available in the near future (Figure 43.5C). The analysis of the EF-hand-containing proteins in *Arabidopsis* also identified other putative enzymes as calcium sensors that could be used in enzymatic reactions using the protein chip platform [6] (Figure 43.5C).

A proteome chip study relevant to calcium signaling was done as a proof of concept for assaying protein chips. A yeast proteome was expressed, purified, and

printed on chips, and the chips were then probed with biotinylated CaM in the presence of calcium [23]. Six known CaM targets were detected along with 33 additional partners including many types that were consistent with a role for CaM. This type of assay should be useful for detecting the targets of the many other calcium sensors, and analysis of the target proteins can lead to identification of domains that bind the sensors. The chips could also be probed with target proteins to identify sensors and activated targets to identify secondary targets involved in a signaling network (Figure 43.5D). A major advantage of the high-throughput protein chips is that complete proteomes can be screened in one assay, and multiple assays changing the concentrations of Ca^{2+} , sensor, or target can be performed yielding not only PPI pairs but also the conditions required for interaction. Overall, whole proteome chips will be useful in identifying (i) all calcium-binding proteins, (ii) targets of calcium sensors, (iii) differences in calcium affinity of sensors, and (iv) global calcium signaling networks.

Analysis of Protein Complexes in Calcium Signaling Using TAP. Y2H interaction studies and the PPI studies using proteome chips are very informative for direct interactions between proteins. However, many components of signaling networks form complexes of proteins that interact directly and indirectly with each other. In order to identify members of complexes, a complete complex needs to be isolated and analyzed. One method developed for this is TAP (Chapters 36 and 37). A cassette consisting of the protein of interest as a fusion to two affinity tags, protein A that binds IgG-sepharose and a CBP that binds CaM-sepharose, is introduced into an organism and expressed at a level equal to its endogenous level. Cell extracts are processed and purified by first using IgG-sepharose, eluting the bound proteins with TEV protease and then binding to CaM-sepharose. Eluted proteins are treated with trypsin, and the resulting peptides can be analyzed by LC-MALDI-TOF-MS and/or nESI-MS/MS. Proteins are then identified by peptide or spectra searches against protein databases such as Mascot (<http://www.matrixscience.com>) or Sequest (<http://www.thermo.com>). This method was used in a high-throughput manner for the yeast proteome. Over 4500 proteins were tagged and 2357 purifications were successful, yielding identification of 4087 proteins organized into 547 protein complexes averaging 4.9 subunits per complex [24].

A great advantage to this method is that it is *in vivo* where the appropriate conditions (i.e., Ca^{2+} levels, presence of all interacting proteins in appropriate concentrations and locations, etc.) are met. This is particularly important for identification of interactions between calcium-sensor targets and their interacting proteins because the activated sensor (or possibly, lack of activated sensor) is necessary for the interaction to occur. The TAP method has been modified for specific uses; in the case of proteins involved in calcium-binding protein networks, the calmodulin-binding protein tag would be inappropriate. In this case the secondary tag could be a His or GST tag and purification would be by anti-His- or GST-sepharose. The calcium sensors and sensor targets identified by bioinformatics, and PPI assays could be used as a subset of proteins to be TAP-tagged and used to isolate complexes for analysis.

Deciphering Calcium Signaling Networks using Protein Chips and Other High-Throughput Methods

The ultimate challenge is to decipher the calcium signaling network—to elucidate all events from a signal to a response and when and how the different components of the signaling pathway are involved. Computational approaches are being developed to analyze the data generated from the proteomics approaches and to integrate data from other high-throughput approaches [25]. The goal of these models is to take qualitative and temporal information and produce a model that explains the relationship between given proteins. These models could organize, display, and simulate the complex protein interaction networks involved in calcium signaling. This type of modeling was used to construct a model of guard cell ABA signaling [26]. In order to model the proteins and their interactions, we need to identify all individual PPIs between calcium sensors and targets and the targets of the sensor targets and their spatial and temporal information. Protein chip assays are a potential high-throughput approach to produce a large body of information for elucidating networks. Protein chips containing potentially all the ORFs for a species can be used in independent assays using a series of proteins and conditions.

43.3 CONCLUSIONS

Constructing the calcium signaling network in plants is not a goal that can be achieved by a protein by protein characterization. The large numbers of proteins involved in the network in primary and secondary interactions and beyond make this a task of overwhelming proportions. However, with the development of high-throughput techniques such as high-throughput Y2H, TAP, MS, and protein chip technology, this goal is becoming achievable. Using these techniques to generate data that can then be analyzed using computational approaches will lead to an understanding of the calcium signaling network.

43.4 FIVE-YEAR VIEWPOINT

Efforts to generate *Arabidopsis* proteome chips are underway. In the next five years or so, an *Arabidopsis* proteome protein chip is expected to be available for the many PPI, small molecule interaction, enzyme, and other functional assays that will be developed as the chip technology develops. This should lead to at least a basic identification of all components in calcium signaling and the function of a majority of the proteins in *Arabidopsis*. Identification of all the calcium sensors, their targets, and the interactome of all calcium signaling proteins should be forthcoming once the protein chips are available for testing. With these data, together with the information obtained with other high-throughput approaches (global expression, localization, genetic, etc.) and new computational tools, elucidation of the basic calcium signaling networks is within reach. From these networks we should be able to identify key hubs in the pathway and generate predictions. Our research efforts will then need to focus on

verifying the pathways that are predicted through the computational modeling of the high-throughput data collected from protein chips and other approaches such as TAP and MS.

REFERENCES

1. Sanders D, et al. 2002. *Plant Cell* **14**:S401–S417.
2. Reddy VS and Reddy AS. 2004. *Phytochemistry* **65**:1745–1776.
3. Reddy ASN. 2001. *Plant Sci* **160**:381–404.
4. Rudd JJ and Franklin-Tong VE. 2001. *New Phytol* **151**:7–33.
5. McCormack E, Tsai Y-C, and Braam J. 2005. *Trends Plant Sci* **10**:383–389.
6. Day IS, et al. 2002. *Genome Biol* **3**:RESEARCH0056, 56.1–56.24.
7. Kolukisaoglu U, et al. 2004. *Plant Physiol* **134**:43–58.
8. Nagata T, et al. 2004. *Mol Biol Evol* **21**:1855–1870.
9. Kung LA and Snyder M. 2006. *Nat Rev Mol Cell Biol* **7**:617–622.
10. Zhu H, Bilgin M, and Snyder M. 2003. *Annu Rev Biochem* **72**:783–812.
11. Fordham-Skelton AP, et al. 1994. *Plant Mol Biol Rep* **12**:355–363.
12. Safadi F, Reddy V and Reddy ASN. 2000. *J Biol Chem* **275**:35457–35470.
13. Reddy ASN, et al. 1996. *Plant J* **10**:9–21.
14. Reddy AS, Reddy VS, and Golovkin M. 2000. *Biochem Biophys Res Commun* **279**:762–769.
15. Reddy VS, Ali GS, and Reddy AS. 2003. *Plant Mol Biol* **52**:143–159.
16. Kim KN, et al. 2000. *Plant Physiol* **124**:1844–1853.
17. Rual JF, et al. 2005. *Nature* **437**:1173–1178.
18. Suter B, Auerbach D, and Stagljar I. 2006. *Biotechniques* **40**:625–644.
19. Rodriguez Milla MA, et al. 2006. *FEBS Lett* **580**:904–911.
20. Ptacek J, et al. 2005. *Nature* **438**:679–684.
21. Oh YH, et al. 2007. *Biosens Bioelectron* **22**:1260–1267.
22. Feilner T, et al. 2005. *Mol Cell Proteomics* **4**:1558–1568.
23. Zhu H, et al. 2001. *Science* **293**:2101–2105.
24. Krogan NJ, et al. 2006. *Nature* **440**:637–643.
25. Schubert P, et al. 2006. *Anal Bioanal Chem* **386**:482–493.
26. Li S, Assmann SM, and Albert R. 2006. *PLoS Biol* **4**:e312.

PART VIII

STRUCTURAL PROTEOMICS

CELL-FREE EXPRESSION SYSTEM FOR EUKARYOTIC PROTEINS

Yaeta Endo and Tatsuya Sawasaki

44.1 INTRODUCTION

Cell-free translation systems can synthesize proteins with high accuracy and speed approaching *in vivo* rates [1], and they can express proteins that seriously interfere with cell physiology. In general, however, they are unstable and thus inefficient [2]. Spirin et al. [3] proposed a continuous-flow cell-free (CFCF) translation method, in which a solution containing amino acids and energy sources is supplied to the translation chamber through a semipermeable membrane. In fact, an *E. coli* cell-free system operated in the CFCF mode gives much higher productivity than a conventional batch system [3]. This system, however, still has inherent limitation because of its prokaryotic nature of translation and folding mechanisms; multidomain proteins, found more often in eukaryotes than in prokaryotes, tend to misfold in prokaryotic systems, whether *in vivo* or *in vitro* [4]. Thus, it seems to be a big challenge to improve prokaryotic cell-free systems for the preparation of multidomain proteins. On the other hand, the rabbit reticulocyte cell-free system, a popular system derived from eukaryotic animal cells, cannot be adopted for the synthesis of preparative amounts of proteins because of its low efficiency and a high endogenous globin content which

obstructs product purification. One of the most convenient and promising eukaryotic cell-free translation systems is conceivably the one based on wheat embryos in which all the components of translation are stored in a dried state, ready for protein synthesis as soon as germination starts. Derived from a eukaryotic plant source, it is expected to have the ability to synthesize eukaryotic multidomain proteins in the folded state. The conventional wheat germ system, however, is plagued with its short life; as a result, it is as inefficient as other cell-free systems.

44.2 DEVELOPMENT OF WHEAT GERM CELL-FREE PROTEIN SYNTHESIS SYSTEM

We have been actively pursuing ways to improve the wheat germ cell-free translation system through its understanding from both fundamental and technological viewpoints. Recently, we found that the conventional wheat germ extracts contain the RNA *N*-glycosidase tritin and other inhibitors of translation such as thionin, ribonucleases, deoxyribonucleases, and proteases which are suspected of being involved in a suicide system targeting at translational machinery. We also ascertained that those inhibitors originate from the endosperm [5, 6]. Extensive washing of wheat embryos could eliminate those endosperm contaminants to produce clean embryos for extracts with a high degree of stability and activity [6]. In fact, this new translation system proved stable not only in its activity and longevity but also in storage; it can be stored for years without loss of activity, even in the lyophilized state. In order to maximize the throughput of the system for practical use, other elemental techniques were developed. They are (i) eliminating both the 5'-mGpppG (cap) and poly(A)-tail (pA) by optimizing the 5'- and 3'-UTRs of mRNA, thereby increasing translation initiation and the stability of the template [7]; (ii) designing the split-primer for PCR to generate transcription templates that can minimize artificial generation of the products [7], a technique that permits high-throughput construction of DNA templates directly from *E. coli* cells carrying cDNAs, with the time-consuming cloning steps bypassed; (iii) constructing an expression vector pEU [7] specialized for the mass production of proteins, which reduces introduction of inevitable mutation during the PCR and abates the cost; (iv) developing the transcription and translation reaction in one tube in which no mRNA purification step is needed [8]; and (v) inventing the bilayer reaction that enables us to perform CFCF mode translation without using a membrane for the high-throughput production [9]. Combining all these elemental techniques, we could establish two cell-free protocols for practical use (Figure 44.1). One of the two is the protocol for materializing genetic information in parallel, which consists of (i) *in silico* selection of suitable genes from the database, (ii) construction of templates for transcription by the split-PCR, (iii) transcription, and (iv) the bilayer translation in which the solution resulting from transcription is directly used as mRNA source. For the production of proteins with such tags as for purification or for the reporter or for both, DNA constructs can be generated by the split-PCR as well. The other protocol is for massive preparation (bold arrow), which consists of (i) selection of suitable gene products from the parallel production described above and subsequent functional

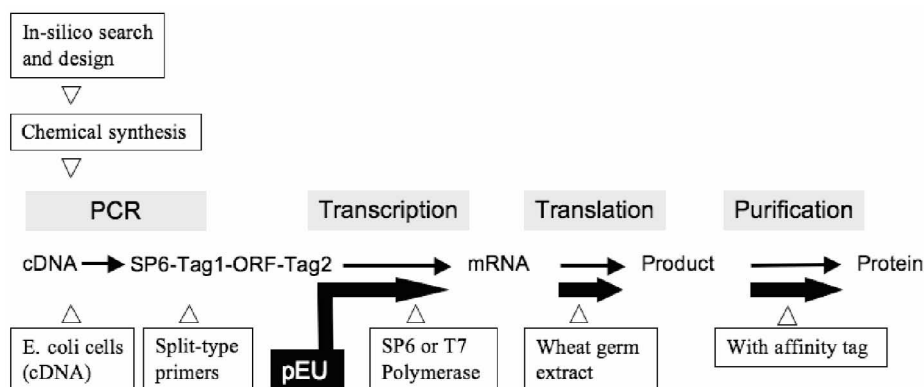


FIGURE 44.1. Procedures for protein synthesis using the wheat germ cell-free system.

screening, (ii) cloning of the genes into pEU and transcription of mRNA, (iii) fine tuning of translational conditions such as the concentrations of pertinent ions, and (iv) protein production that incorporates either the bilayer or dialysis methods. Proteins with a desired purification tag can easily be affinity purified, and the tag portion can be removed by proteolytic cut at a designed linker sequence if desired.

44.3 APPLICATION TO FUNCTIONAL PROTEOMICS

In order to evaluate the performance of the cell-free method in high-throughput screening, we carried out parallel protein syntheses from 27 genes originating with *E. coli* cells carrying the cDNAs [7]. In 50 out of the total 54 cases (authentic and fusion proteins with GST), a clearly visible CBB-stained protein band was obtained. The yield after 36 h of incubation was estimated by densitometric scanning of the bands using BSA as the standard. It was from 0.1 to 2.3 mg/mL of the reaction volume in the dialysis cup mode reaction (CF CF) (Table 44.1). Some gene products were recovered in soluble forms in the supernatant (S) and some others in the precipitate phase (P) after centrifugation at 30,000 g for 15 min. It is worth mentioning here that we could not detect any dependency of the productivity and solubility of the proteins on the gene sources. Furthermore, the system had little preference in codon usage, which is a prerequisite for genome-wide protein expression. In fact, we could express the quality of malaria proteins from 76% AT-rich cDNAs of *P. falciparum* [10]. The most important requirement for an expression system, however, is that it produces a quality product. To verify the quality of proteins and the general applicability of the wheat germ cell-free system, we tested a subset of product proteins for their functional activity (Figure 44.2). PHOT1 gene from *A. thaliana* codes for a protein of 120 kDa which noncovalently binds to flavin mononucleotide (FMN) and presumably acts as a chromophore in light-dependent autophosphorylation [11]. As shown in Figure 44.2A, the protein that was responsible for the blue-light-responsive activity could be produced only when the translation was carried out in the presence of FMN [12]. The result

TABLE 44.1. An Example of Proteins Synthesized in the PCR-Directed Wheat Germ Cell-Free System

Proteins Encoded by cDNAs	Annotation No.	MW	Authentic		Fusion		Clone Name
			Total (mg)	Sup (%)	Total (mg)	Sup (%)	
<i>Arabidopsis (from GenBank and MIPS)</i>							
Chlorophyll a/b-binding protein	X56062	25,995	0.2	30	0.5	90	At01
Agamous-like gene 9 (AGL9)	AF015552	29,065	0.7	30	0.8	90	At02
Flowering locus T (FT)	AF152096	19,808	0.3	100	0.8	100	At03
HY5	AB005456	18,462	0.4	90	1.5	90	At04
Flowering locus F (FLF)	AF116527	21,864	0.2	100	0.4	100	At05
Hypothetical protein (from a commercial cDNA library ^b)	At1 g69630 ^a	11,311	0.1	40	0.1	100	At06
Putative heat shock protein 40	AL021749	38,189	1.8	30	1.0	80	At07
Heat shock protein 70-3	Y17053	71,144	0.9	100	— ^c	— ^c	At08
Putative s-adenosylmethionine synthetase	AY037214	42,793	1.5	100	0.6	100	At09
NADPH thioredoxin reductase	Z23108	40,635	0.1	10	0.5	20	At10
Putative ACC oxidase	AF370155	36,677	1.0	10	1.2	100	At11
Putative fructokinase	AF387001	35,276	1.0	10	0.6	100	At12
Rubisco activase	X14212	51,981	0.4	20	0.6	80	At13
Glutaredoxin	At4 g15660 ^a	11,311	— ^c	— ^c	0.4	80	At14
Chlorophyllase 2	AF134302	34,902	0.2	10	0.4	70	At15
<i>Human (from GenBank)</i>							
Neuron-specific gamma-2 enolase S	M22349	47,266	1.0	100	0.5	100	Hs01
zeta-crystallin/quinone reductase	L13278	35,205	2.3	80	1.3	100	Hs02
X11-like protein	AB014719	82,480	0.5	100	0.2	80	Hs03
Importin alpha 1	NM_002266	57,859	— ^c	— ^c	0.2	30	Hs04
Glyceraldehyde-3-phosphate dehydrogenase	M17851	36,051	0.4	70	0.9	100	Hs05
Enolase 3 ^d	NM_001976	46,956	1.7	80	0.9	100	Hs06
APBA3 ^{d,e}	NM_004886	61,451	0.9	100	0.2	100	Hs07
JAK binding protein ^d (from a commercial cDNA library ^f)	NM_003745	23,550	0.4	30	0.2	100	Hs08
Phosphoglycerate kinase 1	XM_010102	43,965	1.0	100	0.7	100	Hs09
Beta-actin	X00351	41,735	1.3	100	0.3	100	Hs10
Hypothetical protein FLJ10652	XM_006938	41,539	— ^c	— ^c	0.2	10	Hs11
Hypothetical protein FLJ10559	XM_001479	35,237	0.7	50	1.0	70	Hs12

^aMIPS *Arabidopsis thaliana* database MAtDB (<http://mips.gsf.de/proj/plant/jsf/athal/index.jsp>)

^bLambda ZAP II Library, product of Stratagene cat #937010.

^cBelow the detectable level.

^dThese genes were cloned from tissue (heart, brain, kidney, liver, placenta) cDNAs (BioChain Institute, Inc., cat #0516001).

^eAmyloid beta (A4) precursor protein-binding, family A.

^fLambda ZAP II Library, product of stratagene cat #936204.

TABLE 44.2. Difficult Proteins Expressed in the Wheat Germ Cell-Free System

Species	Activity	Gene name	Reference
Cypovirus	Crystalline particle formation	VP3	13
Bacillus	Endonuclease (Bam HI)	Restriction enzyme	9
Thermococcus	DNA synthesis	KOD DNA polymerase	9
Aquifex	Methyltransferase	tRNA (Gm18) methyltransferase	14
Arthrobacter	Sarcosine oxidation	Sarcosine oxidase	9
Arabidopsis	Protein phosphorylation by light	PHOT1	12
Rice	Anthranilate synthase	Anthranilate synthase alpha subunit	15
Rice	Shkimate kinase	Shkimate kinase	16
Rice	EPSP synthase	EPSP synthase	16
Mouse	Immunoglobulin induction	Osteopontin	17
Human	DNA binding	c-fos	18
Human	DNA binding	c-jun	18

demonstrated that a prosthetic group is required during translation and folding into the holoenzyme, which supports the notion that co-translational folding takes place on eukaryotic ribosomes [4]. Figure 44.2B shows the results of the preparation and activity measurement of human kinases. Each of the 11 kinases was synthesized in a GST-fused form, affinity purified with the GST-column, separated by SDS-PAGE, and stained with CBB. When the enzymes were incubated with [γ - 32 P]-ATP in the presence the well-known substrates histone H1 and MBP, subsequent autoradiography clearly demonstrated that each enzyme has its own unique substrate specificity [12]. This suggested that each product might be enzymatically active. Figure 44.2C shows another example in which the polyhedrins of baculoviruses synthesized in the cell-free system were active enough to be assembled *in vitro* into polyhedra [13]. We also attempted to express certain genes whose proteins, if synthesized in *E. coli* cells, exhibited little activity. Our cell-free system proved to be able to produce some of those difficult proteins in active form without the refolding process. Examples of such successful protein synthesis have already been reported in other places and are summarized in Table 44.2. Although the system has not yet been optimized for producing membrane proteins in full size in a soluble, folded state, there is the possibility of combining detergents or liposomes [19–21].

44.4 APPLICATION TO STRUCTURAL PROTEOMICS

One of the bottlenecks in high-throughput structural determination of proteins is the step to produce proteins in folded state. The other limiting step is amino-acid-specific, selective labeling for assigning the signals collected in the ^1H – ^{15}N HSQC spectra in NMR spectroscopy, or selenomethionine labeling for MAD phasing in X-ray crystallography. Cell-free systems have a general advantage over cell expression systems:

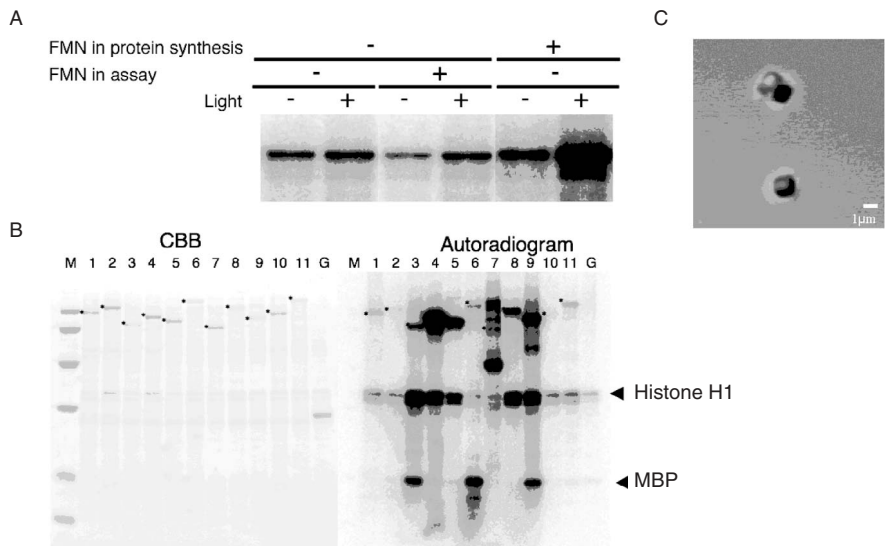


FIGURE 44.2. Quality of eukaryotic gene products expressed in the wheat cell-free system: **(A)** Blue-light-induced autophosphorylation activity of PHOT01 gene product. **(B)** Activity of 11 protein kinases of human beings. Each affinity-purified kinase with GST at N-terminal was treated with Lambda protein-phosphatase and incubated with [γ - 32 P]-ATP in the presence of added histone H1 (from bovine calf thymus) and MBP (from bovine brain). Samples were separated by SDS-PAGE, stained with CBB for analysis (left figure), and subjected to autoradiography (right figure). **(C)** After the translation of polyhedrin gene from baculoviruses in the cell-free system, the translation mixture was kept for 48 h at 4°C. Microscopic observation revealed the formation of small polyhedra.

Cell-free products for NMR measurements do not need extensive purification, because none of the endogenous proteins are labeled during the translation. Cell-free systems derived from *E. coli*, however, contain high levels of amino-acid-metabolizing enzymes, since the extracts are prepared from cells in the exponential growth phase. As a result, many types of inhibitors of amino acid metabolism or even suitable mutant strains have to be added to avoid scrambling of the label among amino acids. In the wheat-based system, in contrast, most of those enzyme activities were expected very low, since the embryos are in hibernation. Nevertheless, we found that two transaminases and one synthetase were active during the translation reaction. Our subsequent search for their inhibitors were successful in that the interconversions of amino- 15 N between Ala and Glu (alanine transaminase) and between Glu and Asp (aspartate transaminase) were completely inhibited by adding β -chloro-L-alanine and L-methionine sulfoximate, and that the other leaking pathway between Glu and Asp (glutamine synthetase) was suppressed by aminooxyacetate [22]. Incorporation of amino acids into protein is not very high even in our system. This economical issue could be solved by recycling the amino acids left in the mixture using the conventional method for purifying amino acids after translation. This is another

attractive feature that benefits the NMR field. The improved methodology may pave the way for a high-throughput protein synthesis system suited for estimating fold-
edness [23, 24] and intermolecular interaction between ligands and the proteins of
inters as well as assigning signals in the NMR field. A fruitful result of the appli-
cation of the wheat cell-free protein synthesis system to NMR structure determina-
tion is now presented at (<http://www.uwstructuralgenomics.org/>) [24, 25]. The results
of X-ray analyses including selenomethionine labeling for MAD phasing are to be
published.

44.5 CONCLUSIONS AND FIVE-YEAR VIEWPOINT

The improved wheat germ cell-free translation system is especially powerful in the
high-throughput production of eukaryotic multidomain proteins in folded state. We
have recently succeeded in the robotic automation of the protocols [8]. The machines
(<http://www.cfsciences.com>) can perform all the steps preceding protein purification,
holding the promise of increasing the throughput and decreasing the cost of protein
production. Protein microarrays, which have enormous potential in biosciences and
medical fields, are one of the promising applications of the cell-free system. Recent
progress of elemental technologies in this field is remarkable [26, 27], but the goal is
still far away. The difficulty is mainly due to the diversity of the biochemical proper-
ties of proteins. One of the most difficult tasks is to establish a technique for storage,
namely, a technique to keep each immobilized protein intact in such a tiny space,
which would be totally different from that for DNA chips. In this light, we would like
to propose possible use of the cell-free system to overcome such difficulty. A prospec-
tive scheme will be to take advantage of the stability and capability of the system
and store all the ingredients involved in transcription and translation in lyophilized
form together with the DNA template. Adding water prior to the preparation of
chips would then start the production of proteins fused with immobilization tags.
Another potential application of the wheat cell-free system in protein chip technology
is for the production of viral polyhedral proteins. Foreign proteins with N-terminal 79
residues (a sequence from BmCPV VP3) can be immobilized in virus-like particles
that resist dehydration and higher temperatures [13]. The accumulated information on
the structures and functions of gene products should revolutionize our understanding of
biology and fundamentally alter the practice in medicine, influencing other industries
as well.

ACKNOWLEDGMENT

This work was partially supported by the Special Coordination Funds for Promoting
Science and Technology by the Ministry of Education, Culture, Sports, Science, and
Technology, Japan.

REFERENCES

Papers of particular interest, published within the annual period of review, have been highlighted as follows: *of special interest, **of outstanding interest.

1. Kurland CG. 1982. *Cell* **28**:201–202.
2. Roberts BE and Paterson BM. 1973. *Proc Natl Acad Sci USA* **70**:2330–2234.
3. Spirin AS, et al. 1988. *Science* **242**:1162–1164.
4. Netzer WJ and Hartl FU. 1977. *Nature* **388**:343–349.
5. Ogasawara T, et al. 1999. *EMBO J* **18**:6522–6531.
6. **Madin K, et al. 2000. *Proc Natl Acad Sci USA* **97**:559–564. The authors clarified a cause of the instability of the cell-free translation reaction, and they developed a method for preparing a highly efficient and yet robust extract from wheat embryos for cell-free protein synthesis.
7. **Sawasaki T, et al. 2002. *Proc Natl Acad Sci USA* **99**:14652–14657. To establish a high-throughput cell-free protein synthesis system, the authors developed a series of essential elementary technologies such as the split-primer PCR and expression vectors specialized for the wheat cell-free system.
8. Sawasaki T, et al. 2007. *Methods Mol Biol* **375**:95–106.
9. Sawasaki T, et al. 2002. *FEBS Lett* **514**:102–105.
10. Tsuboi T, et al. 2008. *Infect Immun* **76**:1702–1708.
11. Sakai T, et al. 2001. *Proc Natl Acad Sci USA* **98**:6969–6974.
12. Sawasaki T, et al. *Phytochemistry* **65**:1549–1555.
13. * Ikeda K, et al. *Proteomics* **6**:54–66. The authors ascertained that polyhedrin synthesized in the wheat cell-free system from baculovirus gene assembled into polyhedra. Using the cell-free system, they found that a sequence of N-terminal 79 residues of BmCPV VP3 peptide was useful for immobilizing foreign protein on the surface of polyhedra. They also found, among other things, that protein immobilized in the polyhedra was protected from dehydration and higher temperatures.
14. Hori H, et al. 2003. *J Biol Chem* **278**:25081–25090.
15. Kanno T, et al. 2004. *Plant Mol Biol* **54**:11–22.
16. Kasai K, et al. 2005. *Planta* **222**:438–447.
17. Miyazaki T, et al. 2005. *Eur J Immunol* **35**:1510–1520.
18. Miyamoto-Sato E, et al. 2003. *Nucleic Acids Res* **31**:e78
19. Klammt C, et al. 2004. *Eur J Biochem* **271**:568–580.
20. Pornillos O, et al. 2005. *Science* **310**:1950–1953.
21. Nozawa A, et al. 2007. *Plant Cell Physiol* **48**:1815–1820.
22. Morita EH, et al. 2004. *J Biomol NMR* **30**:37–45.
23. Morita EH, et al. 2003. *Protein Sci* **12**:1216–1221.
24. ** Vinarov DA, et al. 2004. *Nat Methods* **1**:149–153. The authors described a wheat germ cell-free platform for protein production that supports efficient NMR structural studies of eukaryotic proteins and offers advantages over cell-based methods.

25. *Vinarov DA and Markley JL. 2005. *Expert Rev Proteomics* **2**:49–55. The authors reviewed development of an automated platform for NMR-based structural proteomics that employs wheat germ extract for cell-free production of labeled protein.
26. Zhu H and Snyder M. 2003. *Curr Opin Chem Biol* **7**:55–63.
27. Merkel JS, et al. 2005. *Curr Opin Biotechnol* **16**:447–452.

PROTEIN STRUCTURE DETERMINATION

Jian-Hua Zhao and Hsuan-Liang Liu

45.1 INTRODUCTION

Structural proteomics, the determination and prediction of the 3D structures of proteins on a genome-wide scale for better understanding their structure–function relationships, has provided a new rationale for structural biology and has now become a major initiative in biotechnology [1]. Structural biologists believe that the whole characters of living cells possess their own molecular and cellular functions, which could be tightly regulated by the 3D structures of proteins encoded in a cell. Consequently, the structural genomics effort, which aims for a structure-based functional discovery of all unannotated gene products, has become a challenging and ambitious research topic.

The 3D structure of a protein could undoubtedly provide an essential clue for the fundamental questions about its biological functions. Structural proteomics projects could fill up the universe of protein folding space and provide an accurate prediction of the structure-function relationship. For structural proteomics research, two major experimental techniques (X-ray crystallography and NMR) spectroscopy have played a central role in determining the structures of proteins on a genome-wide scale.

Unfortunately, these techniques are expensive and time-consuming. On the other hand, sequencing of proteins is relatively fast, simple, and inexpensive. Thus, there is a large gap between the number of known protein sequences and the number of known 3D structures. This gap has grown over the past decade (and is expected to keep growing) as a result of the completeness of various genome projects worldwide. Therefore, computational methods are becoming increasingly important because they may give some indication of structure and/or function of proteins. Since it was found that proteins can fold into their unique native state without any additional genetic mechanisms, over 25 years of effort has been expended on the determination of the 3D structure from the sequence alone, without further experimental data. Despite the amount of effort, the protein folding problem remains largely unsolved and is thus one of the most fundamental problems in computational molecular biology today. There are three major approaches to solve this problem: homology modeling, protein threading, and *ab initio* prediction.

This chapter introduces recent developments in structural proteomics for protein structure determination, including (a) instrumental methods such as X-ray crystallography and NMR spectroscopy and (b) computational methods such as homology modeling, protein threading, and *ab initio* structure prediction.

45.2 INSTRUMENTAL METHODS FOR PROTEIN STRUCTURE DETERMINATION

X-Ray Crystallography

Introduction. To date, X-ray crystallography is still the most powerful technique for structure determination and analysis of proteins because it is capable of providing atomic coordinates of the whole assembly [2, 3]. Ever since the determination of the structure of myoglobin, these structural data have contributed tremendously to our understanding of structural biology at the atomic level [4]. It is particularly useful in determining the protein structure with PTMs, which is frequently found in a living organism. In addition, the molecular sizes of proteins determined by X-ray crystallography with modern synchrotron facility range from a small protein up to the large 70S ribosomal complex (~2500 kDa), which could cover a whole protein universe in a cell [4]. However, there are currently several bottlenecks that must be overcome to achieve the goal of high-throughput. These include the expression, purification, and crystallization of proteins and the structure solving process.

Protein Structure Determination by X-Ray Crystallography.

Expression and Purification. The first step in obtaining a novel protein structure is the generation of sufficient amounts of highly purified protein for crystallization trials; typically, 10–50 mg of protein is required at this stage. Almost all proteins subjected to structural studies are expressed in heterologous systems. Recent developments in molecular biology have allowed high-throughput cloning and expression of proteins. However, overexpression in host cells may result in inappropriate PTM and incorrect

folding, which usually forms inclusion bodies and insoluble aggregates and thus must be discarded. Such limitations have been overcome by a number of strategies, such as using genes from different species, altering constructs, screening for solubility, and utilizing different cellular or cell-free expression systems. Furthermore, several novel refolding techniques, such as refolding chromatography based on molecular chaperones that are part of the cellular machinery responsible for folding nascent protein [5] and the designed small-molecule agents to assist the refolding process [6], have been developed to recover these proteins. If successful, these methods could have impact on improving the rate of protein production.

Crystallization. Once the proteins are correctly expressed and purified, the next step is to form crystals of sufficient quality to collect high-resolution data for structure determination. Crystallization is usually regarded as a slow, resource-intensive step with low success rates in obtaining good quality of crystals. However, much of the failure in this step can be attributed to poor protein quality. The crystallization step normally commences after the successful preparation of a highly purified and soluble sample [7]. Therefore, the use of biophysical methods such as dynamic light scattering to assess the quality of a protein is a key step before performing crystallization experiments [8]. The actual crystallization conditions can at present not be determined *ab initio*, and the standard protocol is to screen and optimize a wide range of conditions, such as pH, ionic strength, temperature, salt and protein concentrations, and cofactors, in this step. Particularly for proteins with low yields, the ability to screen more conditions at the required protein concentration becomes critical. In other words, there are many unpredictable variables and multiple conditions that should be tried to get an optimal crystal for each protein.

Until now, crystallization has benefited enormously from automation and technologies allowing the use of small protein sample volumes over the past few years [9]. Meanwhile, crystallization robots have contributed to increasing the efficiency of screening crystallization conditions. Usually, conventional crystallization of protein required relative high protein concentrations in solution (about 10 mg/mL). Therefore, the ability to handle a small amount of solution, such as the nano-screening technique, greatly speeds up the crystallization procedure. Additionally, these robots also provide image capture and analysis systems in place, which monitor the drops automatically and look for the presence of crystals using edge-detection methods [9].

Data Collection. Once the crystallization step is completed, the crystals are prepared for the diffraction data collection. The diffraction data arises from the coherent interference of all molecules in the crystal. The diffraction data comprise a radially distributed array of diffraction maxima whose positions are determined by the lattice repeat and any internal symmetry found in the crystal. The data can be most easily visualized as a series of spots that are uniformly spaced but have different intensities. The variation of the intensities of the spots arises from the interference of scattered X rays from atom within a single molecule and as a consequence contains information on the 3D arrangement of the atoms within the protein molecule.

This information can then be used to visualize the scattering matter, electrons, by calculating an electron density map. The data used in this calculation include the

intensities measured directly from the crystal along with phases that go into a Fourier summation. Although the analysis of diffraction data starts after the data have been recorded, the experiment and the analysis are tightly linked together. Data collection at record speed is now feasible on third-generation synchrotron sources; with the use of CCD detectors (The Woodlands, TX), a data set can be collected within minutes.

Solving the Phase Problem. To construct an electron density map into which a macromolecular model can be built, both the amplitudes and the phases of the diffracted X rays need to be known. While the amplitudes can be straightforwardly derived from the intensities, the phases have to be obtained indirectly. During the last few decades, X-ray crystallographic methods have focused on improving the measurements and recovering this phase information. Several methods have been developed to solve the phase problem—for example, the purely computational *ab initio* methods if extremely high-resolution X-ray data are available [10]; molecular replacement [11] if a homologous structure is known; or heavy atom substitution techniques. Additionally, recent advances in molecular biology coupled with the availability of tuneable synchrotron radiation are the main driving forces for the success of the MAD technique [12]. The ability to analyze metalloproteins by this method, or to analyze protein crystals that have been soaked in solutions containing heavy atoms or halides, have popularized the MAD and single-wavelength (SAD) measurements [13]. This technology provided an efficient route to solving the crystallographic phase problem but is crucially dependent on obtaining the experimental data with highest quality.

Model Building and Refinement. Once there is an atomic model inside the density map, the next step is to understand it locally. Thus, the problem of fitting an initial model is to see the forest for the trees. To get an overview of a density map, it needs to be simplified in some way. Generally, it is done through skeletonization, or making a bones representation of map. In skeletonization, the map is reduced to a series of lines that run between the peaks of density through regions of reasonably high density. With this representation, a lot of information can be presented quite simply. Then, one can start to fit the polypeptide chain by choosing plausible α -carbon positions from the bones. After the main chain is traced, one has to fit the side chains. The macromolecular model is then subjected to a refinement procedure in which the parameters of the model are optimized to best fit the experimental data and stereochemical expectations. During the refinement, the electron density map can be improved, and this may require considerable adjustment and rebuilding of the model.

The Accuracy of the Structure. The resolution of an X-ray structure determination is a measure of how much data are collected. The greater the amount of data, the more detailed the features in the electron-density map to be fitted, and, of course, the greater the ratio of the number of observations to the number of atomic coordinates and B-factors to be determined. Resolution is expressed in angstroms, with a lower number signifying a higher resolution. Some favorable protein crystals usually permit data collection to ultrahigh resolution: 1.5 Å and below.

Conclusions. The technological advances in X-ray crystallography over the past decade have resulted in the rapid growth of protein structure data. However, crystal structures of multicomponent systems and membrane proteins are still limited, providing key challenges for these new approaches. However, it is highly foreseen that these novel approaches will continue to generate protein crystal structures at unprecedented rates and drive a new wave of structure-based drug discovery.

NMR Spectroscopy

Introduction. Although X-ray crystallography is regarded as the potential workhorse for structural proteomics, the throughput of structure determination using it remains unclear. In contrast, NMR spectroscopy does not require crystals and thus samples appropriate for structure determination can be identified in a relatively short period of time. Moreover, NMR experiments can be carried out in aqueous solution under conditions similar to the physiological conditions in which the protein normally functions. These particular features have provided insights into the structure–function relationships for a large number of proteins. However, structure determination by NMR is currently limited by size constraints and lengthy data collection and analysis times. Nevertheless, NMR spectroscopy still plays a significant role in structural proteomics even with its current limitations [14].

Protein Structure Determination by NMR Spectroscopy

Protein Solution. The first step in this strategy is the preparation of protein solution. A suitable sample, usually about 0.5 ml of 1 mM protein solution, is carefully prepared. If the molecular mass of a protein is larger than 10 kDa, enrichment with ^{13}C and ^{15}N isotopes is required to resolve spectral overlap in ^1H NMR spectroscopy. ^{13}C and ^{15}N are used because the most abundant carbon isotope (^{12}C) does not give a NMR signal and the most abundant nitrogen isotope (^{14}N) has desired NMR properties. Fully isotope labeled proteins can be produced by growing the bacteria on minimal medium containing $^{15}\text{NH}_4\text{Cl}$ and $^{13}\text{C}_6\text{-glucose}$ as sole N and C sources, respectively. In principle, all necessary NMR measurements could be done with one [^{15}N , ^{13}C]-labeled sample.

Processing of NMR Spectra. This sample is subsequently used to record a set of multidimensional NMR experiments, which provide the NMR spectra after suitable data processing. At this stage, the chemical shift plays a very important yet mainly indirect role in structure determination by NMR. The chemical shift is very sensitive to the microenvironment of a particular nucleus. Therefore, in a folded protein, multiple copies of the same amino acid can be distinguished as a result of the conformation-dependent chemical shift. However, a direct interpretation of the chemical shift is usually not (yet) possible. For protein spectra, this is only possible in multidimensional NMR spectra that are measured on NMR instruments operating with strong magnetic fields corresponding to a proton resonance frequency of 600 MHz or higher.

Sequential Resonance Assignment. For a detailed analysis of the information content of NMR spectra, nearly complete assignments of signals in the spectra to individual atoms in the molecule are a prerequisite. All procedures use the known protein sequence to connect nuclei of amino acid residues which are neighbors in the sequence (i.e., the assignment procedure takes advantage of the sequential arrangement of the individual amino acids in a polypeptide chain).

The first experiment is the [^1H , ^1H]-COSY, which detects through-bond interactions between protons and correlates protons that are separated by up to three chemical shift bonds. With the experiment the protons within an amino acid can be correlated. However, neighboring amino acids in the polypeptide sequence cannot be connected. For this purpose, the second experiment, the [^1H , ^1H]-NOESY [15–17], is measured. Resonances in this spectrum manifest short distances between pairs of protons. Only short distances with backbone amide protons are extracted at this stage and used to establish correlations between protons of neighboring residues in the polypeptide sequence. With this information the spin systems can be put in their proper sequential order. A comparison of the sequentially ordered spin systems with the known amino acid sequence allows one to assign the sequence position to every spin system and thus complete sequence-specific resonance assignments are obtained [18].

Collection of Conformation Constraints. Generally, NMR data alone would not be sufficient to determine the positions of all atoms in a biological macromolecule. It must be supplemented by information about the covalent structure of the protein. Although a variety of NMR parameters contain structural information, the crucial information comes from NOE measurements which provide distance information between pairs of protons. The quality of a solution structure increases with the number of consistent input constraints used in the structure calculation. In other words, higher accuracy of the structure can be obtained with more number of constraints. The quality of a solution structure determined by the number of consistent input constraints used in the structure calculation. For a high-resolution structure determination, the maximal possible number of NOE constraints must be collected as input for the calculation of the complete 3D protein structure. The NOE input data have the format of allowed distance ranges.

Calculation of the 3D Structure. Mainly two approaches of computer programs are used to calculate the 3D structures based on NMR constraints. The first approach uses interatomic distance [19, 20] and the second one work with torsion angles of chemical bonds [21, 22]. These programs calculate in the end the Cartesian coordinates of spatial molecular structures which are consistent with the set of constraints obtained from NMR data. The experimental constraints do not uniquely describe one exact 3D structure because NMR-derived constraints typically describe a range of possible values and many distances cannot be determined. Therefore, the structure calculation is repeated many times to determine an ensemble of structures consistent with the input data set. A good ensemble of structures minimizes violations of input constraints and samples the complete conformational space allowed by the constraints. For this reason

the NMR structures are usually represented by a bundle of structures. The quality of a structure is assessed by the calculation of the root mean square deviation (RMSD) between the atoms of individual conformers in the bundle.

Recent Technological Advances. The foundation of successful NMR studies is the high-quality spectra with good sensitivity and resolution. When studying larger molecules in solution, these basic requirements become more difficult to fulfill. Thus the corresponding spectra show poor resolution and sensitivity. In practice, it becomes a key challenge to determine structures from proteins which have MWs above 40 kDa. Important advances in extending this size limit have recently been made with the introduction of novel NMR techniques and new biochemical approaches [23]. A novel spectroscopic concept, transverse relaxation optimized spectroscopy (TROSY) based on selection of slowly relaxing NMR transitions, also can provide significant sensitivity enhancement for large proteins [23, 24]. Furthermore, the TROSY experiments enable the recording of high-quality NMR spectra of macromolecules and supermolecular structures with MWs above 100 kDa [25].

The second challenge for NMR spectroscopy is the reduction of the data collection time required for structure determination. Besides, the development of NMR experiments that allow matching of instrument time investments to the minimum time required for measuring the chemical shift data is another challenge. They can be overcome by reduced dimensionality experiments [26, 27], with simultaneous frequency labeling of more than one atom type in indirect dimensions. Another advance involves partial deuteration [28], which provides samples that can be studied with improved S/Ns that result from their sharper linewidths and longer transverse relaxation times.

These technologies together provide the basis for high-throughput NMR and are particularly valuable for samples with limited stabilities and low solubilities. Another important issue of development involves automated analysis of NMR data. The design of cryogenic probes will certainly assist in reducing the bottleneck of time required for data collection in NMR-based structural proteomics. In addition, some automated resonance- and nuclear Overhauser effects-assignment programs, such as *NOAH* [29], *AUTOASSIGN* [30], *ARIA* [31], *ANSIG* [32], *TATAPRO* [33], and *SANE* [34], hold great promise in reducing the amount of time in structure determination. These recent developments provide automated analysis of NMR assignments and the 3D structures of proteins up to 200 amino acids [35]. However, general methods for automated analysis of side-chain resonance assignment are not yet well developed, and there are as yet no examples of completely automated protein structure determinations.

Conclusions. NMR plays a key role in several established pilot projects for protein structure determination. The primary challenges for high-throughput applications of NMR are the long time periods for data collection and the laborious expert reasoning required for data analysis. Recent advances in probe design, data collection strategies, and software engineering demonstrate the potential for higher-throughput data collection and automated structure analysis.

Comparisons Between X-Ray Crystallography and NMR

Unit now, NMR remains plagued by heavy instrumentation requirements, the need to interpret the data manually, the size limitation of ~ 250 residues, and the difficulty of assessing the statistical correctness of the structure. However, NMR is on the verge of a dramatic increase in efficiency, with improvements on instrumentation and automation of data analysis. On other hand, protein crystallography, which does not have size limitations, is also progressive and may become an almost fully automated process from crystal to structure. Therefore, crystallography may dominate the first stages of the structural genomics efforts. However, it may turn out to be a serious bottleneck because of the incapability to predict and control protein crystallization. In this regard, the advantages of NMR are that it does not require excessive protein purity and it avoids the requirement to grow crystals. Eventually, it can be predicted that the structural genomics projects could evolve to the point where many small proteins will be undertaken by NMR and crystallography will be the method of choice for larger proteins and protein complexes.

45.3 COMPUTATIONAL APPROACHES FOR PROTEIN STRUCTURE PREDICTION

Homology Modeling

Introduction. Model-building by homology is a useful technique when one wants to predict the structure of a target protein of known sequence that is related to another protein of known sequence and structure. If the two proteins are closely related, the known protein structure will form the basis for a model of the target. Although the quality of the model depends on the degree of similarity of amino acid sequence, it is possible to predict the quality from sequence. Homology modeling is based on two major observations. First, the structure of a protein is uniquely determined by its amino acid sequence [36]. Knowing the sequence should, at least in theory, suffice to obtain the structure. Second, during evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structure and distantly related sequences still fold into similar structures. This relationship was first identified by Chothia and Lesk [37] and later quantified by Sander and Schneider [38]. Homology methods usually follow four basic steps: (i) searching templates, (ii) sequence alignment, (iii) model building, and (iv) model validation (Figure 45.1).

Protein Structure Prediction by Homology Modeling

Searching Templates. Homology modeling usually starts by searching the PDB of known protein structure using the target sequence as the query [39]. This search is generally carried out by comparing the target sequence with the sequence of each of the structures in the database. Template search and alignment are essential for the correctness and the quality of a homology model. Homology modeling programs

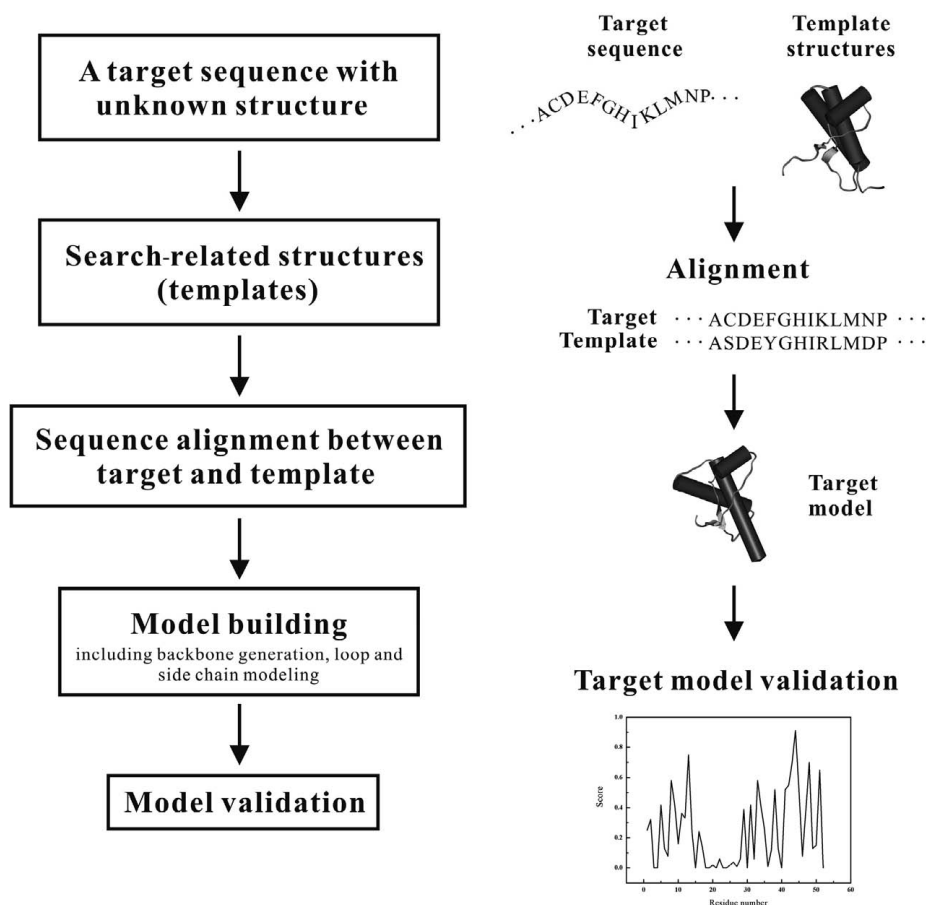


FIGURE 45.1. Steps in homology modeling.

always generate a structure for any target sequence using the conformation of the template structures and the alignments between the target protein sequence and its templates. If the templates or the alignments are incorrect, the output model will certainly be wrong as well. Empirically, if there is no determined structure with at least 25% sequence similarity to the target sequence, then there may be no template available that is suitable for reliable homology modeling. Thus, the potential accuracy of a homology model is positively related to the percentage of sequence identity on which it is based. High-accuracy homology models are based on more than 50% sequence identity to their templates. They tend to have about 1-Å RMSD for the main-chain atoms, which is comparative to the accuracy of a medium-resolution NMR structure or a low-resolution X-ray structure [40].

There are two main categories of protein comparison methods. The first category compares the target sequence with each of the database sequence independently, using pairwise sequence-sequence comparison [41]. *FASTA* [42] and *BLAST* [43] are the

most-well-known programs in this category. The second category depends on multiple sequence comparisons to improve the sensitivity of search [43–47]. This approach begins by finding all sequences in a sequence database that are clearly related to and easily aligned with the target. The multiple alignment of these sequences is the target sequence profile, which implicitly carries additive information about the location and pattern of evolutionarily conserved positions of the protein. The most popular program in this category includes *PSI-BLAST* [43], which implements a heuristic search algorithm for short motifs. A further step to increase the sensitivity of this approach is to precalculate sequence profiles for all known structures and then use a pairwise dynamic programming algorithm to compare the two profile. This has been implemented in *PDB-BLAST* [47]. The construction of profile-based HMM is another sensitive way to locate universally conserved motifs among sequences [48].

Sequence Alignment. Once a list of potential templates is obtained using searching methods, it is necessary to select one or more templates that are suitable for the particular modeling problem. Then, one can start to align the amino acid sequences of two proteins. It will generally be observed that insertions and deletions lie in the loop regions between helices and sheets, especially if the proteins are closed-related. For all regions not involving insertions or deletions, the main chain of known structure provides the model for the corresponding residues of the target structure. Alignments can be improved by including structural information from the template. For example, gaps should be avoided in secondary structure elements, in buried regions, or between two residues that are far apart in space. In addition, it is not necessary to select only one template. In fact, the use of several templates is generally another approach to increase the model accuracy [49].

Model Building.

BACKBONE GENERATION. Once the alignment is ready, the actual model building can be start. Creating the backbone is trivial for most of the model: Simply copy the coordinates of those template residues that show up in the alignment with the model sequence. If two aligned residues differ, only the backbone coordinates can be copied. If they are the same, one can also include the side chain.

LOOP MODELING. The accuracy of loop modeling is a major factor in determining the usefulness of comparative methods. However, loops are generally too short to provide sufficient information about their local fold because the environment of each loop is uniquely defined by solvent and the protein that cradles it. There are two main approaches to loop modeling: (i) *knowledge-based*: searching homologues loop structures using a database of loop fragments where the loop must be a good geometric fit to the region of the target protein [37, 50] and (ii) *energy-based*: as in true *ab initio* fold prediction, using energy function to judge the quality of loop, followed by minimization of the function using Monte Carlo [51] or molecular dynamics techniques [52] to arrive at the best loop conformation. Often, the energy function is modified to facilitate the search [53].

SIDE-CHAIN MODELING. Practically all successful approaches to side-chain placement are at least partly knowledge-based. They use libraries of common side-chain conformation (rotamers) extracted from high-resolution X-ray structures. The various rotamers are tried successively and stored with a variety of energy functions. Because of the combinatorial explosion (the choice of a certain rotamer automatically affects the rotamers of all neighboring residues, which in turn affect their neighbors, and so on), one might intuitively expect rotamer prediction to be computationally demanding. However, the number of combinations is in fact so large that even nature could not try all of them during the folding process, indicating that there must be methods developed to reduce the search space. The good news is that certain backbone conformations strongly favor certain rotamers (allowing, for example, a hydrogen bond between side chain and backbone) and thus shrink down the search space. Some position-specific rotamers libraries are widely used today [54–56]. To predict a rotamer, the corresponding backbone stretch in the template is superposed on all collected examples and possible side-chain conformations are selected from the best backbone matches [57].

Model Validation. After a model is built, it is important to check for possible errors. The accuracy of a model can be approximately predicted from the sequence similarity between the target and the template. If it is greater than 90%, the accuracy of the model can be compared to crystallographically determined structure, except for few individual side chains [58, 59]. From 50% to 90% identity, the root mean square error in the modeled coordinates can be as large as 1.5 Å, with considerably larger local errors. If the sequence identity drops to 25%, the alignment turns out to be the main bottleneck for homology modeling, often leading to very large errors. Generally, the errors in comparative models can be divided into five different categories: (i) errors in side chain packing; (ii) distortions or shifts of a region that is aligned correctly with the template structures; (iii) distortions or shifts of a region that does not have an equivalent segment in any of the template structures; (iv) distortions or shifts of a region that is aligned incorrectly with the template structure; and (v) a misfolded structure resulting from using an incorrect template.

Errors in a model become less of a problem if they can be localized. An essential step in the homology modeling process is therefore the verification of the model. There are two ways to estimate errors in a structure: (i) calculation the model's energy based on a force field: this method checks whether the bond lengths and bond angles are within normal ranges and whether there are lots of bad contact in the model (corresponding to a high van der Waals energy); and (ii) determination of normality indices that describe how well a given characteristic of the model resembles the same characteristic in real structures. Many features of protein structures are well-suited for normality analysis. These include: general checks for the normality of bond lengths, bond and torsion angles [60, 61], inside/outside distributions of polar and apolar residues [62], and the distance and direction of atomic contacts [63]. Most methods used for the verification of models can also be applied to experimental structures (and hence to the templates used for model building). A detailed verification is essential

when trying to obtain new information from the model, either to interpret or predict experimental results or to plan new experiments

Conclusions. The recent advances in homology modeling—especially in detecting distant homologues, aligning sequences with template structures, modeling of loops and side chains, and detecting errors in a model—have contributed to reliable prediction of protein structure, which was not possible even several years ago. It can be expected that homology modeling will steadily increase its importance because it helps to bridge the gap between the available sequence and structure information by providing reliable and accurate protein models and will also contribute to a better understanding of the structure–function relationships of proteins.

Protein Threading

Introduction. Methods for protein structure prediction offer some hope for narrowing the gap between the number of known protein sequences and the number of experimentally determined protein structures. Threading (fold recognition) methods can be very effective despite their slowness, the requirement for human intervention to interpret the results, and the inaccuracy of sequence–structure alignments produced. Threading methods were originally developed to recognize pairs of proteins, which have no obvious similarities in sequence but have similar folds. Recently, it has become clear that in fold recognition it is useful to distinguish between pairs of proteins that are homologous and those that are analogous. Where no evolutionary relationship is believed to exist between two structurally similar proteins, threading would be the only applicable method for identifying this type of relationship.

Typically, algorithms comparing protein sequence have been used to recognize close and distant homologies and to define homologous families. Because they are based entirely on sequence similarity, it is widely accepted that proteins with statistically significant sequence similarity are homologous. Such algorithms could be used to compare proteins with unknown structure. By contrast, threading algorithms have been used to predict protein structure and, when introduced, are based heavily on structural features even if the sequence similarity is extremely low or nonexistent.

Protein threading (sequence–structure alignment) [64–68] is a promising template-based method for fold recognition, which identifies a appropriate fold from a structure library for the query sequence and provides an alignment between the query protein and the fold. The basic idea of threading can be summarized as follows (see Figure 45.2): First, one has a target sequence and a library of templates or known structures. Presumably, these are protein data bank structures and the library contains all known protein folds. Next, one takes the target sequence and “threads” it through each template in the library. The word *threading* implies that one drags the sequence step by step through each location on each template, but really one is searching for the best arrangement of the sequence as measured by some score or quasi-energy function. Finally, all the candidate models with their scores are collected. The best scoring (lowest energy) one is then taken as the predicted structure. A threading requires four components: (i) the library of representative protein folds for use as

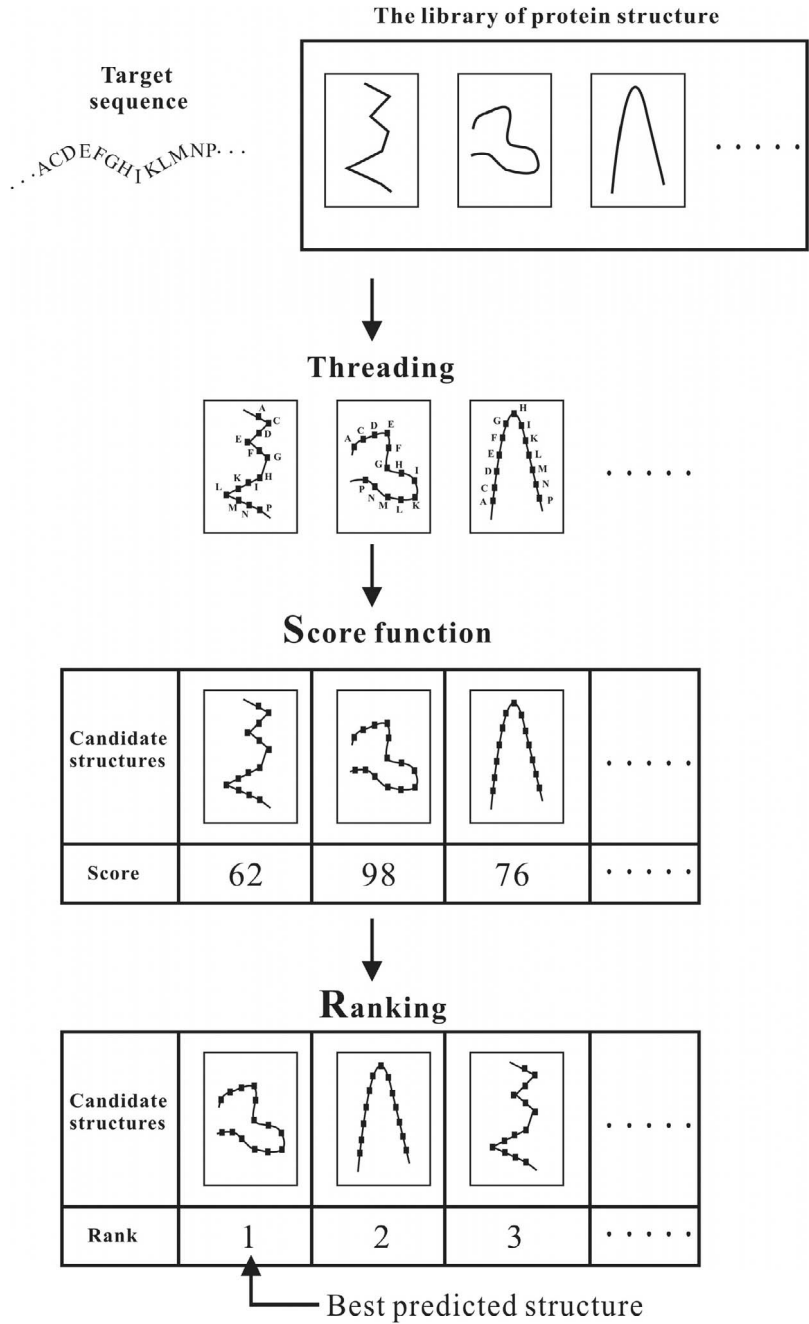


FIGURE 45.2. Steps in protein threading.

templates; (ii) an energy function to describe the fitness of any alignment between target and template; (iii) a threading algorithm to search for the lowest energy among the possible alignments for a given target–template pair; and (iv) a criterion to estimate the confidence level of the predicted structure.

Protein Structure Prediction by Threading Methods.

The Library of Protein Folds. From a protein family, one should select the member that has the best-quality coordinates or the one that is in some way most representative of the family of structures. Continuing in this vein, the idea of representative structures implies that one has already clustered all known proteins down to a set of families. This could be based on sequence identity or some measure of structural similarity. Moreover, one is not even limited to simple PDB coordinates. Madej et al. [69] used a library based on extracted cores from proteins, and some have suggested that the structures in the library could be optimized so as to make the ranking of models statistically more reliable [70].

The Energy Function. Threading score (energy) functions are also usually more coarse-grained than those used in a real energy calculation. In a threading calculation, the sequence residues are placed on the backbone of the template structure; from there, one can calculate the ideal coordinates for the C_β atom. A threading score function usually represents each residue by one or a few interaction sites. Often, most of the chemical identity of a residue comes from an interaction site located at the C_β atom or a point closer to the side-chain center of mass. With this level of representation, it is not common to depend only on pure physics. For example, a threading score function does not usually have a term like Coulombs law for electrostatics or a Lennard-Jones term for other atomic interactions. Instead, there is a common approach to build a score function: potentials of mean force. Potentials of mean force derived from the statistical analysis of interaction regularities in proteins can reliably recognize grossly misfolded structures or wrong crystallographic models [71] and can assess the quality of models prepared in homology modeling [72]. And of course, the same potentials can be used in fold recognition.

The Threading Algorithm. The threading approach can be further subdivided into two classes: (1) *Threading that considers only the preference of amino acids in the target sequence at single sites of the templates (singleton threading).* Singleton threading creates a 1D structure profile for each residue position in a template structure using local 3D environmental information such as secondary structure type, degree of environmental polarity, and the fraction of the residue surface accessible to solvent. The energy function is based on the compatibility of the 20 amino acids for each position in the 1D structure profile. The compatibility is derived from the statistics of the whole template database. Optimal 1D alignment between a target sequence and a template can be determined by dynamic programming. The final template is selected according to the optimal score or its statistical significance. The singleton threading can combine secondary structure predictions and position-dependent profiles based on

multiple sequence alignments into the energy function. Several servers are available for singleton threading: *123D* [73], *TOPITS* [74], *SAS* [75], and the *UCLA-DOE Structure Prediction Server* [76]. (2) *Threading that uses the preference on pairs of amino acids in the target sequence within a contact distance when they are aligned to a given structure.* Threading using pairwise interactions considers the propensity of two amino acids in the target sequence to be aligned within a specified distance using a score function compiled from a database of structures. Several threading programs using pairwise interactions are available, including *NCBI Threading Packag* [77], *PROFIT* [65], *PROSPECT* [67], and *THREADER* [66].

The threading approach is more sensitive than the sequence-based search methods such as *PSI-BLAST* [78] and *SAM-T98* [45, 48]. However, the main difficulty for threading is that the structure profile and the residue pair derived from the template may not adequately describe the corresponding information in the target protein as a results of the structure difference between the two proteins, even when they share the same fold. This is a more significant problem in the fold category than in the superfamily category.

The Criterion to Estimate the Confidence Level. Once a model is built, it is important to evaluate the confidence of the predicted structure. A model should be consistent with experimental observations, such as site-directed mutagenesis, cross-linking data, and ligand binding.

Conclusions. Since threading has already changed since the early implementations, it is also clear that the methods will continue to evolve. Some techniques combine elements of threading with methods for *ab initio* structure predictions are now available [79–82]. It holds the promise of being able to predict structures unlike any previously being solved. Threading may also be applied in new contexts such as macromolecular interactions and multimolecular assemblies [83, 84].

Ab Initio Methods

Introduction. The ultimate goal of protein structure prediction is to predict the structure of a protein without any reference to know protein structure. Unlike comparative methods that are limited to protein families with at least one known structure, *ab initio* prediction is free of such a restriction. In addition, if fold recognition methods are not able to find a suitable template structure to build a model for a gene product, in an ideal world it would be possible to apply an *ab initio* method to build an approximate model. The *ab initio* methods aim to predict the structure from sequence alone [85, 86]. The *ab initio* methods assume that the native structure corresponds to the global free energy minimum accessible during the lifespan of the protein and attempt to find this minimum by an exploration of many conceivable protein conformations.

There are two components to *ab initio* prediction: (i) devising a search method to explore the conformation space and (ii) a potential (or energy) function that can distinguish between the correct (native or native-like has lower energy) structure from the incorrect ones. In many methods, the two components are coupled together such

that a search method drives, and is driven by, the scoring function to find the native or native-like structure.

Protein Structure Prediction by Ab Initio Methods.

Search Methods. To accelerate conformational search, one must employ techniques that permit coarse sampling of the energy landscape. A variety of methods may be used in conjunction with reduced complexity models and simplified potentials to perform broad searches through low resolution structures, including Metropolis Monte Carlo simulated annealing [87], simulated tempering [88], evolutionary algorithms [89], and genetic algorithms [90].

A single search is unlikely to find the global minimum of the free energy landscape and may instead yield a structure that has become trapped in a local minimum. In an effort to correct for this possibility, many current methods perform numerous conformational searches to generate an ensemble of candidate structures. Numerous techniques have now been used to select those structures most likely to be close to the native structure from the ensemble [91–93]. Further insights into features of native protein structures and properties of near-native ensembles will undoubtedly add to the arsenal of methods of selecting the most native-like structures.

Potential Functions. There are two categories of potentials that may be employed in evaluating the free energy of the peptide chain and the surrounding solvent. Firstly, molecular mechanics potentials seek to model the forces that determine protein conformation using physically based functional forms parameterized from small molecule data or *in vacuo* quantum mechanical (QM) calculations. Secondly, protein structure-derived potentials or scoring functions are empirically derived from experimental structures from the PDB [94, 95]. Usually a functional form is not specified and instead pseudoenergies are obtained by taking the logarithm of probability distribution functions. Such structure-derived potentials are particularly useful in conjunction with reduced complexity models, where they may be viewed as representing the interactions between.

Both classes of potentials must represent the forces that determine macromolecular conformation: solvation, electrostatic interactions including hydrogen bonds and ion pairs, van der Waals interactions, and, in certain cases, covalent bonds [96]. Additionally, they must be applicable at a granularity that is in keeping with that of the representation selected and the target resolution of the method.

A Developed Method: Rosetta Predictions. *Ab initio* tertiary structure prediction from sequence has proven to be extremely difficult even after tremendous efforts in the past decades [90, 97–101]. *Ab initio* prediction programs require long computing time and the prediction results are generally unreliable. However, some recent developments using hierarchic approaches, which first build local structures and then assemble them into a global structure, seem to provide new hope for generating low-resolution structures. Once local structures are more or less defined, assembling them requires

a significantly smaller computational search space. The optimization process is typically carried out using genetic algorithms [102] or Monte Carlo simulations [99]. Local structures can be built through fragment libraries.

Recent advance in this area introduces a toolkit, ROSETTA, for *ab initio* protein structure prediction [51]. ROSETTA is based on the assumption that the distribution of conformations sampled for a given nine residue segment of the chain is reasonably well approximated by the distribution of structures adopted by the sequence in known protein structures [103] (see Figure 45.3). Fragment libraries for all possible three- and nine-residue segments of the chain are extracted from the protein structure database by using a sequence profile comparison method. The conformational space defined by these fragments is searched using a Monte Carlo procedure with an energy function that favors compact structures with paired β -strands and buried hydrophobic residues. The output structures are by construction consistent with the local conformational biases inherent in the sequence and have low free-energy nonlocal interactions by virtue of the Monte Carlo optimization procedure. For each query sequence a large number of independent simulations are carried out starting from different random number seeds. The resulting structures are clustered and the centers of the largest clusters are selected as the highest confident models

This strategy resolves some of the typical problems with both the conformational search and the free energy function. The search is greatly accelerated because switching between different possible local structures can occur in a single Monte Carlo step, and less demand are placed on the free energy function since local interactions are accounted for in the fragment libraries.

45.4 CONCLUSIONS

According to the widely accepted “thermodynamic hypothesis,” the native conformation of a protein corresponds to a global free energy minimum of the protein/solvent system [104]. Therefore, having a correct energy function, one could use the tools of computational physics to search for the native structure in the conformational space. However, this approach is still unable to reliably predict a previously unknown structure of a protein for which only a sequence is known. Two principal problems for *ab initio* prediction of protein structure are the lack of adequate molecular potentials and the enormous size of the conformational space of even the smallest protein. Even so, recently, there have been significant advancements in the field. There is hope that *ab initio* methods will continue to improve and that this improvement will provide both fundamental insights into the physics underlying protein folding and a valuable and practical resource for genome analysis.

45.5 FIVE-YEAR VIEWPOINT

In the post-structural genomics era, a functional analysis of the unknown gene products would be one of most important issues in understanding the basic mechanism of

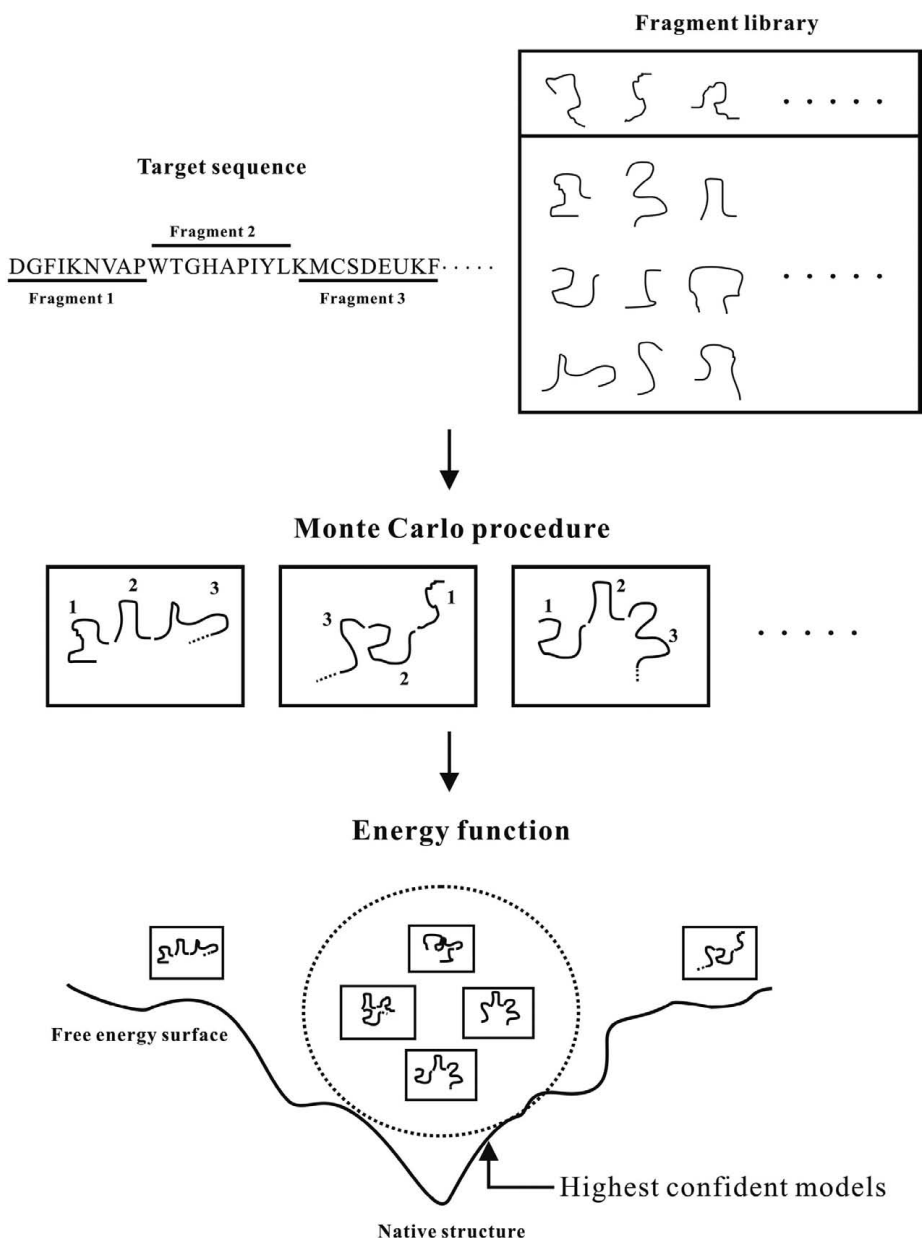


FIGURE 45.3. Steps in Rosetta prediction.

living cells. The structural proteomics is a new research area to provide us a complete understanding of organisms based on structural information of a single protein or the assembly of protein complexes. From a practical viewpoint, the structural information derived from structural genomics will be directly used for drug development in the pharmaceutical industry. The advantages of the X-ray crystallography and NMR spectroscopy are very useful in structural proteomics research. A combination or integration of these methods would be mandatory to fill up a complete dictionary of protein folding space on a genomic scale. Therefore, these two methods will continue to play synergistic roles in structural proteomics. On the other hand, structure prediction methods will be a crucial tool for the generation of the full 3D structure maps. As the number of structural templates increases, a vast experimental pool of data will be available for future machine learning experiments and hopefully lead to a rapid improvement in the accuracy and speed of such efforts. Furthermore, establishing protein interaction networks is crucial for understanding cellular operations. The 3D structures can be used to interrogate the whole interaction network to validate and infer molecular details for the interactions proposed by other approaches. We can expect that the determination and prediction of the 3D structures and the structure–function relationships of proteins for studying PPIs will become increasingly important in the post-genomic era.

REFERENCES

1. Smith T. 2000. *Nat Struct Biol Suppl* **7**:927.
2. Ban N, et al. 2000. *Science* **289**:905–920.
3. Zhang GY, et al. 1999. *Cell* **98**:811–824.
4. Sali A, et al. 2003. *Nature* **422**:216–225.
5. Altamirano MM, et al. 1997. *Proc Natl Acad Sci USA* **94**:3576–3578.
6. Vuillard L, Rabilloud T, and Goldberg ME. 1998. *Eur J Biochem* **256**:128–135.
7. Stevens RC. 2000. *Structure* **8**:R177–R185.
8. Moreno A, et al. 2000. *J Mol Struct* **519**:243–256.
9. Abola E, et al. 2000. *Nat Struct Biol* **7**:973–977.
10. Uson I and Sheldrick GM. 1999. *Curr Opin Struct Biol* **9**:643–648.
11. Turkenburg JP and Dodson EJ. 1996. *Curr Opin Struct Biol* **6**:604–610.
12. Ogata CM. 1998. *Nat Struct Biol* **5**:638–640.
13. Dauter Z, Dauter M, and Rajashankar KR. 2000. *Acta Crystallogr D Biol Crystallogr* **56**:232–237.
14. Christendat D, et al. 2000. *Nat Struct Biol* **7**:903–909.
15. Jeener J, et al. 1979. *J Chem Phys* **71**:4546–4553.
16. Kumar A, Ernst RR, and Wüthrich K. 1980. *Biochem Biophys Res Com* **95**:1–6.
17. Wider G, et al. 1984. *J Magn Reson* **56**:207–234.
18. Wüthrich K. 1986. *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, New York.
19. Havel TF. 1991. *Prog Biophys Mol Biol* **56**:43–78.
20. Nilges M, Clore GM, and Gronenborn AM. 1988. *FEBS Lett* **229**:317–324.

21. Güntert P, Braun W, and Wüthrich K. 1991. *J Mol Biol* **217**:517–530.
22. Güntert P, Mumenthaler C, and Wüthrich K. 1997. *J Mol Biol* **273**:283–298.
23. Wider G and Wüthrich K. 1999. *Cur Opin Struct Biol* **9**:594–601.
24. Wuthrich K. 1998. *Nat Struct Biol* **5**:492–495.
25. Salzmann M, et al. 2000. *J Am Chem Soc* **122**:7543–7548.
26. Szyperski T, et al. 1993. *J Am Chem Soc* **115**:9307–9308.
27. Szyperski T, et al. 1998. *J Biomol NMR* **11**:387–405.
28. Gardner KH and Kay LE. 1998. *Annu Rev Biophys Biomol Struct* **27**:357–406.
29. Mumenthaler C and Braun W. 1995. *J Mol Biol* **254**:465–480.
30. Zimmerman D, et al. 1997. *J Mol Biol* **269**:592–610.
31. Nilges M and O'Donoghue S. 1998. *Prog Nucl Magn Reson Spectrosc* **32**:107–139.
32. Helgstrand M, et al. 2000. *J Biomol NMR* **18**:329–336.
33. Atreya H, et al. 2000. *J Biomol NMR* **17**:125–136.
34. Duggan BM, et al. 2001. *J Biomol NMR* **19**:321–329.
35. Moseley HNB and Montelione GT. 1999. *Curr Opin Struct Biol* **9**:635–642.
36. Epstein CJ, Goldberger RF, and Anfinsen CB. 1963. *Cold Spring Harbor Symp Quant Biol* **28**:439–449.
37. Chothia C and Lesk AM. 1987. *J Mol Biol* **196**:901–917.
38. Sander C and Schneider R. 1991. *Proteins* **9**:56–68.
39. Bourne PE, et al. 2004. *Nucleic Acids Res* **32**:D223–D225.
40. Baker D and Sali A. 2001. *Science* **294**:93–96.
41. Apostolico A and Giancarlo R. 1998. *J Comput Biol* **5**:173–196.
42. Pearson WR. 2000. *Methods Mol Biol* **132**:185–219.
43. Altschul SF, et al. 1997. *Nucleic Acids Res* **25**:3389–3402.
44. Henikoff JG, et al. 2000. *Electrophoresis* **21**:1700–1706.
45. Krogh A, et al. 1994. *J Mol Biol* **235**:1501–1531.
46. Gribskov M and Veretnik S. 1996. *Methods Enzymol* **266**:198–212.
47. Jaroszewski L, Pawlowski K, and Godzik A. 1998. *J Mol Model* **4**:294–309.
48. Karplus K, Barrett C, and Hughey R. 1998. *Bioinformatics* **14**:846–856.
49. Sanchez R and Sali A. 1997. *Proteins* **29**:50–58.
50. Jone TA and Thirup S. 1986. *EMBO J* **5**:819–822.
51. Simons KT, Bonneau R, Ruczinski I and Baker D. 1999. *Proteins* **37**:171–176.
52. Fiser A, Do RK, and Šali A. 2000. *Protein Sci* **9**:1753–1773.
53. Tappura K. 2001. *Proteins* **44**:167–179.
54. de Filippis V, Sander C, and Vriend G. 1994. *Protein Eng* **7**:1203–1208.
55. Stites WE, Meeker AK, and Shortle D. 1994. *J Mol Biol* **235**:27–32.
56. Dunbrack RL Jr and Karplus M. 1994. *Nat Struct Biol* **5**:334–340.
57. China G, et al. 1995. *Proteins* **23**:415–21.
58. Chothia C and Lesk AM. 1986. *EMBO J* **5**:823–36.
59. Sippl MJ. 1993. *Proteins* **17**:355–362.
60. Morris AL, et al. 1992. *Proteins* **12**:345–364.
61. Czaplewski C, et al. 2000. *Protein Sci* **9**:1235–1245.

62. Baumann G, Frommel C, and Sander C. 1989. *Protein Eng* **2**:329–34.
63. Vriend G and Sander C. 1993. *J Appl Crystallogr* **26**:47–60.
64. Bowie JU, Luthy R, and Eisenberg D. 1991. *Science* **253**:164–170.
65. Sippl MJ and Weitckus S. 1992. *Proteins* **13**:258–271.
66. Jones DT, Taylor WR, and Thornton JM. 1992. *Nature* **358**:86–89.
67. Xu Y, Xu D, and Uberbacher EC. 1998. *J Comp Biol* **5**:597–614.
68. Crawford OH. 1999. *Bioinformatics* **15**:66–71.
69. Madej T, Gibrat JF, and Bryant SH. 1995. *Proteins* **23**:356–369.
70. Huber T and Torda AE. 2002. Protein structure prediction by threading, Force field philosophy, approaches to alignment. In: *Protein Structure Prediction, A Bioinformatic Approach* (Tsigelny IF, ed.), International University Line, La Jolla, CA, pp. 263–298.
71. Luethy R, Bowie JU, and Eisenberg D. 1992. *Nature* **356**:83–85.
72. Jaroszewski L, et al. 1998. *Protein Sci* **7**:1431–1440.
73. Alexandrov NN, Nussinov R, and Zimmer RM. 1996. In: *Biocomputing, Proceedings of the 1996 Pacific Symposium Singapore* (Hunter L and Klein T, eds.), World Scientific Publishing Co., Singapore, pp. 53–72.
74. Rost, B. 1995. TOPITS: Threading one-dimensional predictions into three-dimensional structures. In: *Third International Conference on Intelligent Systems for Molecular Biology* (Rawlings C, et al., eds.), AAAI Press, Cambridge, England, pp. 314–321.
75. Milburn D, Laskowski RA, and Thornton JM. 1998. *Protein Eng* **11**:855–859.
76. Fischer D and Eisenberg D. 1996. *Protein Sci* **5**:947–955.
77. Bryant SH and Lawrence CE. 1993. *Proteins* **16**:92–112.
78. Park J, et al. 1998. *J Mol Biol* **284**:1201–1210.
79. Jones DT. 2001. *Proteins* **45**:127–132.
80. Skolnick J, et al. 2001. *Proteins* **45**:149–156.
81. Zhang Y, Kolinski A, and Skolnick J. 2003. *Biophys J* **85**:1145–1164.
82. Kihara D, et al. 2001. *Proc Natl Acad Sci USA* **98**:10125–10130.
83. Lu L, Lu H, and Skolnick J. 2002. *Proteins* **49**:350–364.
84. Lu L, et al. 2003. *Genome Res* **13**:1146–1154.
85. Bonneau R and Baker D. 2001. *Annu Rev Biophys Biomol Struct* **30**:173–189.
86. Pillardy J, et al. 2001. *Proc Natl Acad Sci USA* **98**:2329–2333.
87. Simons KT, et al. 1997. *J Mol Biol* **268**:209–225.
88. Hansmann UH and Okamoto Y. 1997. *J Comput Chem* **18**:920–933.
89. Bowie JU and Eisenberg D. 1994. *Proc Natl Acad Sci USA* **91**:4436–4440.
90. Pedersen JT and Moulton J. 1997. *J Mol Biol* **269**:240–259.
91. Park B and Levitt M. 1996. *J Mol Biol* **258**:367–392.
92. Huang ES, et al. 1996. *J Mol Biol* **257**:716–725.
93. Samudrala R and Moulton J. 1998. *J Mol Biol* **275**:895–916.
94. Sippl MJ. 1995. *Curr Opin Struct Biol* **5**:229–35.
95. Koppensteiner WA and Sippl MJ. 1998. *Biochemistry (Mosc)* **63**:247–252.
96. Park BH, Huang ES, and Levitt M. 1997. *J Mol Biol* **266**:831–846.
97. Li Z and Scheraga HA. 1987. *Proc Natl Acad Sci USA* **84**:6611–6615.
98. Friedrichs MS and Wolynes PG. 1989. *Science* **246**:371–373.

99. Skolnick J and Kolinski A. 1991. *J Mol Biol* **221**:499–531.
100. Sali A, Shakhnovich E, and Karplus M. 1994. *J Mol Biol* **235**:1614–1636.
101. Unger R and Moult J. 1992. *J Mol Biol* **5**:637–645.
102. Simons KT, Strauss C, and Baker D. 2001. *J Mol Biol* **306**:1191–1199.
103. Anfinsen CB. 1973. *Science* **181**:223–30.
104. Privalov PL and Gill SJ. 1988. *Adv Protein Chem* **39**:191–235.

PART IX

OTHER TOPICS IN PLANT PROTEOMICS

PROTEOMICS IN CONTEXT OF SYSTEMS BIOLOGY

Serhiy Souchelnytskyi

46.1 INTRODUCTION

Why Proteomics Requires Systems Biology

Any biological system such as a body, tissue, cells, and organelles consists of a number of components that interact with each other and change their properties upon such interactions. Interactions between components may be physical or functional—for example, formation of a complex or an indirect influence of one component on functions of another, respectively [1]. This is valid for cells originating from plantae, animalia, fungi, protista, and monera kingdoms. Proteins are components that are essential for any cellular function, because they represent enzymatic activities, structural components, and regulatory signals. Comprehensive studies of proteomes become possible with development of proteomics, which allows (a) description of thousands of proteins and their PTMs and (b) evaluation of protein activities in given cells at a given time point [2–5]. This generates huge volumes of information that is rarely fully used for analysis of biological models. To explore large-scale data generated in a single proteomics experiment by traditional so-called in-depth studies would require work of hundreds of researches during two to five years, with the rate of successful

analysis of less than 30 %. Thus, such approach would be inefficient. To select proteins crucial for a studied model, researchers should be able to analyze proteomics data in a fast and comprehensive way. Tools to do such an analysis are available in the systems biology field.

Proteomics generates data that increasingly fulfill requirements of systems biology for quantitative and qualitative information [6, 7]. Description of thousands of proteins in various intracellular compartments and upon a defined time period becomes more and more routine in proteomics. In this review, recent developments that pave the way for the integration of proteomics and systems biology are discussed. These developments include the standardization of data acquisition and the increased comprehensiveness of proteomics studies in description of functional status, localization, and dynamics of proteins. In the context of systems biology, modeling tools become adapted to analysis of proteomics data (Figure 46.1). This indicates that proteomics is reaching the level when it is not only a “fishing tool” to select a particular protein for a single-protein study, but also a technique for comprehensive overview of systemic effects in biological models. Systemic effects are defined as effects which can be observed only as a result of interactions between components of a model (Table 46.1). Thus, from the one side we observe adaptation of systems biology tools to specifics of proteomics, and from the other side proteomics provides data that are suitable for a large-scale analysis. In this chapter we will analyze what systems biology requires, what proteomics can deliver, and what their integration may look like.

46.2 WHAT SYSTEMS BIOLOGY REQUIRES

Comprehensiveness

Systemic properties can be observed only when a certain number of components, or species, is analyzed. The number of species can vary from two, as required for a simple feedback loop, to hundreds, as required for analysis of complex processes (e.g., cell proliferation or differentiation). Thus, the definition of a required number of species in order to observe systemic effects is not strict. However, comprehensiveness of analysis is a feature that can be described as “more is better”: The more the species of the studied model are explored, the better the description of this model will be.

Practically, selection of species for a systemic study is a balance between increasing a number of species for the most comprehensive description and decreasing a number of studied species due to experimental limitations. This selection is based on the nature of studied biological phenomena and on a mathematically defined minimal number of components which is required to observe systemic behavior of an abstract system. Biological phenomena include number of reactions and components, which have to be considered for an informative modeling. On the other side, abstractive modeling has shown that certain systemic effects can be observed only if a defined number of species are involved. As an example, a toggle switch requires three connected components [8–10]. Generally, proteomics provides description of a sufficient number of species to observe systemic effects. However, dependencies between these species are not always known, which is limiting for unveiling systemic properties.

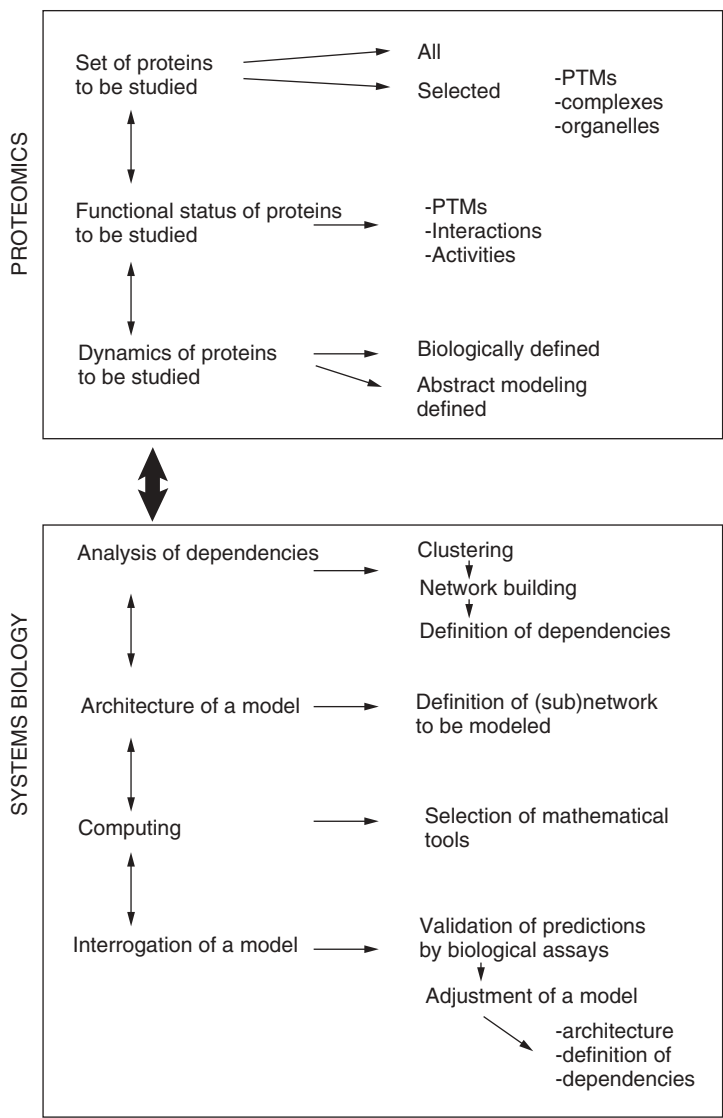


FIGURE 46.1. Proteomics and systems biology are shown as two blocks with description of their specific approaches and types of data. Although an integration of proteomics and systems biology is a reiterative process with constant optimization of all components, the figure indicates a general flow of such an experiment. Selection of studied proteins (top of the figure) defines a model, and proteomics tools provide data sets of a certain type of comprehensiveness and details about species/proteins in their dynamics. Various clustering approaches lead to building of networks and description of dependencies between species/proteins. This defines the most optimal architecture of a model which is further computed. Experimental interrogation of a model (bottom of the figure) provides proofs supporting a model.

TABLE 46.1. Definition of Terms

Term	Definition
Proteome	Protein complement of genome
Proteomics	Technologies to study proteome in a comprehensive large-scale way
Systems biology	Study of systemic properties of biological systems
Model	The description of biological processes in mathematical terms, without discrimination of mathematic tools
Modeling	The application of methods to analyze complex, real-world problems in order to make predictions about what might happen with various actions

Details about Species

The second requirement for species is information about their quantity and quality. For proteins, it includes information about expression levels, functional status, and localization of proteins. Or in the other words, how much, how active, and where proteins are. This challenge is dealt with in two ways. The first way is to employ large-scale technologies to collect all required information at once. Although technologies to do so are available, it would require coordinated actions of various research groups representing required expertise in measurements of protein expression, activities, and localization. Thus, there is no a single study that was able to collect such comprehensive information for one experimental model in a single experiment.

The second way is to define a model to be analyzed in such a way that required information about species would be provided by available experimental tools. For example, modeling of CNK1-dependent Erk1/2 activation upon TGF β treatment includes seven key species that were monitored by immunoprecipitation and immunoblotting techniques and would not require information about localization of proteins [11]. Immunoblotting and immunoprecipitation would be sufficient to evaluate protein quantities and activities of interest. Although these techniques may not provide information about absolute quantities of proteins, for the modeling purpose, information about relative quantities may be sufficient. A validation of the prediction would be performed by transcriptional assays, which is an example of an independent technique to prove the model [11]. Thus, an informative modeling can be performed with limited details about species and using techniques available in a single laboratory.

Details about species may also define which mathematical tools would be most suitable for modeling [8–10, 12]. For example, abundance in detailed information—for example, with absolute quantities and activities of proteins—may be required for informative modeling using a system of ordinary differential equations (ODE). On the other side, if there is no description of direct dependencies between species, statistical correlation analysis tools would be the most suitable for modeling [8–10, 12]. Thus, requirements for details about species can vary, although the level of details is defined by the coverage of a biological model and on selected modeling tools.

Dynamics

The third requirement is a description of changes of quantity, functional status, and localization of species. Knowledge of proteome dynamics is essential for selection of mathematical tools that would be used for modeling. This is also important for definition of a time period over which the model would be studied [8–10]. Collection of information about dynamics of species may be based on empirical knowledge of a studied biological model and on knowledge of behavior of abstract systems. For example, to explore activation of Erk kinase, one would have to measure protein changes every minute over a period of the first 15 min, and then every 2 or 5 min over a period of up to 2 h [13]. Dynamics would have to be measured for expression and/or phosphorylation of mitogenic ligands, their receptors proteins, MAPK cascade, and proteins that modulate Erk activation.

Solving the problem of optimization of data collection intervals would improve the design of experiments. An example of such an approach has been reported by Wolkenhauer and colleagues [14], who applied a multiple-shooting method to a system of ordinary differential equations describing a simple dynamic model. A possibility of calculating the frequency and number of sampling points would strongly enhance the efficiency of proteomics studies, because it will warrant collection of data at time points significant for modeling and will capture all the crucial dynamic features of a proteome. The concept for the optimization of sampling time has been developed for simple dynamic systems, and its application to complex systems (e.g., proteomes), is the next challenge. Thus, combination of experimental knowledge of a studied system and tools for optimization for data collection points should be able to deliver information about dynamics of a proteome.

46.3 WHAT PROTEOMICS PROVIDES

Current proteomics techniques allow us to evaluate how much of a particular protein is present in cells or other samples, what is its functional status, in which compartment of a cell a protein is located, and how fast this protein changes its expression level and properties. In this section, we will analyze how this information fulfills requirements of systems biology.

Protein Expression

An evaluation of protein expression includes detection, identification, and quantification of proteins. Simultaneous detection of hundreds of proteins in an intact cell is still not realistic due to technical limitations. Therefore, disruptive techniques (i.e., protein extraction), are employed to obtain proteins in stabilized solutions. Plant cells and tissues contain significant amounts of compounds that interfere with protein separation—for example, metabolites, lipids, and polysaccharides. Thus, extracted proteins often have to be separated from contaminants, which is achieved by precipitation of proteins by TCA, acetone, or phenol extraction followed by precipitation

[reviewed in references 2 and 3]. Solubilization of precipitated proteins is an important step, because it has to provide maximal recovery of proteins.

Protein extracts are then subjected to separation, with 2-DGE and LC being the most often used techniques. It is estimated that these techniques may allow analysis of up to 50 % of a cellular proteome [4, 5]. Enrichment techniques may enhance coverage of proteome analysis, but only for a selected group of proteins. For example, use of anti-pTyr antibodies allows studying of proteins phosphorylated on Tyr residues [15]. As another example, IMAC and lectins allow enrichment for phosphorylated or glycosylated proteins [16–18]. Chemical modification of phosphorylated or glycosylated amino acid residues is another way to enrich proteins with these PTMs and have been described in recent publications [19–21]. The SELDI method is based on a selective binding of proteins to particular surfaces—for example, ion-exchanger, reverse-phase, or other capturing matrices. Thus, SELDI can be considered as an enrichment technique. Promising is the combination of comprehensive and unbiased 2-DGE and LC with enrichment of proteins of particular groups, because it may provide a high degree of proteome coverage.

Protein chips have been developed to enrich and measure expression of selected proteins [17]. Various protein-capturing agents have been described—for example, antibodies, aptamers, and proteins, or their domains. Protein chips are useful for monitoring selected proteins, but novel nonannotated proteins would not be detected. Protein chips are known as expression, interaction, and structural arrays [17]. Chips are important for the interrogation of any models that analyze protein expression, PPIs, and protein modifications. Although protein chips have limited value for protein separation, they allow fast and reliable analysis of a relatively large number of components.

Peptide-based proteomics is gaining a significant attention, due to its high capabilities of peptide separation by MS coupled with LC [4]. The essence of this approach is that peptides generated after digestion or chemical cleavage of proteins are entities that are analyzed, assuming that peptides represent proteins. On the other side, an approximation of detection of few peptides to a full-length protein may not be justified, especially in the case of shotgun proteomics. In shotgun proteomics, proteins are not separated before generation of peptides, and the same peptide may originate from various post-translationally modified forms of the same protein [4]; for example, a full-length and proteolytic fragment of a protein may produce the same peptides.

Separated proteins have to be detected. Detection techniques include staining of protein spots, spectral analysis, and MS analysis. The most often used staining techniques include CBB, silver, and fluorescence dye staining. These techniques are well established and were reviewed elsewhere (Chapter 2).

Proteins can also be labeled before separation, which increases sensitivity and/or provides detection of a selected group of proteins. Fluorescent probes have been efficiently used in DIGE to detect proteins [22]. It has to be noted that improvements of DIGE labeling chemistry resulted in an efficient and robust technique. Fluorescent dyes can also be used to detect phosphorylated or glycosylated proteins—for example, Pro-Q DPS and Pro-Q EPS, respectively [23, 24].

Detected proteins have to be identified. Today there is no alternative to MS in efficiency of identification. PMF and various peptide fragmentation techniques provide identification of any kind of protein in quantities down to attomole [2–5]. This is significantly lower than the identification levels of other techniques—for example, the Edman degradation method. Moreover, MS allows identification of PTMs in the same samples that were used for identification of proteins.

For identification of particular proteins or groups of proteins, we can use specific antibodies, lectins, or protein domains. Such identification is limited to selected proteins, and may require additional validation experiments which would make the whole technique too cumbersome to be used in a large-scale study. Thus, MS is currently the technique of choice for identification of proteins, and recent developments of MS were reviewed elsewhere (Chapter 3).

Quantitative presentation of protein species is essential for informative modeling. 2-DGE-based proteomics allow quantification by staining of protein spots. Fluorescent probes, CBB or silver staining give acceptable evaluation of a relative expression of proteins [reviewed in reference 5]. Staining allows generally good, though variable, linearity and sensitivity of protein detection and quantification. Modified labeling of proteins with fluorescence dyes in DIGE allows evaluation of relative levels of proteins in samples that are mixed after labeling with different fluorescent dyes, and they are run in one gel. Staining is an efficient way to visualize proteins, but it provides assessment of a relative, and not an absolute, protein concentration.

The relative expression of proteins can also be estimated by comparison of selected peptides using ICAT, iTRAQ, and SILAC and ICPL techniques [25–28]. The combined fractional diagonal chromatography (COFRADIC) approach can also be used for evaluation of protein expression [29]. However, peptide/MS-based proteomics is not suitable for an evaluation of absolute quantities of proteins [5, 15, 30]. Peptide/MS-based proteomics can be an acceptable method of evaluating the relative expression of proteins under conditions that expression levels of full-length proteins are confirmed.

Saturation metabolic labeling of cells with ^{35}S -labeled amino acids was proposed to estimate the absolute quantities of proteins [31, 32]. The quantity of proteins would be calculated from the absolute radioactivity incorporated in the proteins and would be corrected to the level of specific radioactivity of a probe and a number of Cys and Met residues in a protein. This technique also allows the evaluation of synthesis and turnover rates of cellular proteins [31, 32].

The above-mentioned methods to evaluate protein expression have been successfully used in studies of human, animals, and plant samples. These methods provide information that is of sufficient quality to fulfill requirements for systemic studies. However, successful modeling would benefit from more comprehensive coverage of proteomes and absolute quantification. These two directions are currently pursued very actively.

Functional Status of Proteins

The functional status of a protein is defined by its folding, PTMs, and interacting molecules, and it defines how this protein will affect a modeled system. Approxi-

mately 200 PTMs create an astonishing variability of protein species [4, 5, 33–35]. Comprehensive simultaneous analysis of all PTMs in a proteome is unrealistic due to technical limitations. However, methods of studying phosphorylation, glycosylation, acetylation, ubiquitylation, sumoylation, and methylation have been developed. Because these PTMs are involved in crucial regulatory processes in cells, there is a strong interest in their study. As an example, phosphorylation on Ser, Thr, and Tyr residues are essential for regulation of cell proliferation, differentiation, migration, and apoptosis. PTMs also trigger PPIs [36, 37].

Phosphoproteins can be enriched by immunoprecipitation with specific Abs and by IMAC [2–5, 15, 16, 26, 27]. Phosphorylation has also been studied by metabolic labeling of proteins with radioactive [^{32}P]orthophosphate, followed by 2-DGE [38, 39]. Phosphorylations on Ser, Thr, and Tyr residues have been studied using β -elimination-directed cleavage at the site of modification, using fluorescent dyes, and using chemical modification of phosphorylated residues [19–21, 40].

MS-based proteomics has proven to be an efficient tool for identifying modified peptides; phosphorylated peptides may be enriched with specific antibodies or by IMAC in the same way as proteins, before being analyzed by MS [2–5]. Glycosylated proteins have been successfully enriched with lectins [41]. The β -elimination followed by Michael addition of dithiothreitol (BEMAD) technique allows identification of *O*-glycosylation sites in purified proteins [19]. Ubiquitylation is an important triggering mechanism for protein degradation, and more than a thousand proteins have been identified using pull-down of His-tagged Ub in *S. cerevisiae* [42]. The analysis of SUMO-2 interacting proteins identified eight novel targets of sumoylation [43]. Thus, tools to study phosphorylation, glycosylation, ubiquitylation, and sumoylation have been developed.

Acetylation and methylation are two PTMs that are important for intracellular signaling. No comprehensive studies of acetylation and methylation have been reported, but the availability of specific antibodies to enrich acetylated or methylated proteins opens up a possibility of such research [33, 44, 45].

Multiplexed use of fluorescence probes that recognize specific PTMs is an interesting approach to the comprehensive analysis of protein phosphorylation, glycosylation, and expression [46]. It is especially attractive, because this technique allows analysis of two different PTMs in the same gel by using two different probes. Many proteins have various PTMs simultaneously, with one PTM affecting another modification. Therefore, the development of techniques for simultaneous studying various PTMs is of great importance and has to be intensified.

Proteins execute their functions in transient or stable complexes with other molecules [47]. Purification of protein complexes, followed by an identification of their components, has been performed on a large scale in a number of studies. Thousands of interactions have been described when a large number of proteins were used as baits [48–50]. These studies have contributed to the definition of connections between protein species, which are essential for design of model architecture.

Two-hybrid and phage display techniques also contribute to building PPI networks [51–53]. These techniques have been widely applied to animal and human proteins, and can be applied to plant proteins.

Comparison of data sets obtained with various techniques has shown that these data sets complement each other, and none of the techniques has provided full coverage of PPIs. As an example, Y2H assays would identify mostly binary interactions, while pull-down assays would describe predominantly protein complexes [47, 54]. PPIs have various affinities and can be stable or transient. Some techniques allow detection of transient interactions (e.g., Y2H), while others will detect only stable complexes. The latter can be exemplified by the TAP, which detects only complexes that are sufficiently stable to be preserved after two sequential precipitation procedures [48]. Thus, considerations of binary interactions *versus* protein complex detection, localization of proteins, and differences in affinities of interactions strongly require combination of PPI data obtained by different techniques. Such combined databases would allow validation of interactions across data sets obtained in various cells, tissues, and species.

Proteomes of every cell type are unique. This suggests an importance of selection of a model system for studying PPIs. As an example, PPIs identified in yeast may not be observed in plant cells, and vice versa. As conditions for interactions, folding and PTMs pattern of the same protein may differ in various types of cells, acquisition of PPI data has to consider used species (plant, yeast, worm, fly, frog, or mammalian cells), and type of tissues.

Protein complexes may interact with each other. Knowledge of complexes that functionally interact via the exchange of identified proteins may lead to the understanding of the functional importance of these proteins. It would define the hubs and connections in functional networks [55]. Identification of a series of interactions within one complex or module, along with parallel connections of a protein to different complexes in PPI networks, defined the dynamics of complexes and unveiled proteins important for the architecture of distribution and the multiplication of signaling [55].

An analysis of intracellular organelles is another way to approach PPIs. A number of protein complexes have been described in plant organelles: mitochondria (complex I, V, mitochondrial ribosomal complex), chloroplast (PS I and II antenna proteins, chloroplast ribosomal complexes), vacuoles, peroxisomes, cell wall, apoplast, nuclei, and membranes [2, 3, 56–58]. The definition of protein composition of organelles does not indicate whether interactions between proteins are direct. This uncertainty of the protein–protein connections in organelles is a problem for the modeling of networks of protein interactions. On the other hand, the composition of organelles provides valuable information about functional complexes and their regulation and therefore contributes to the definition of functional clusters.

Direct assessment of protein activities has been approached by ABPP, which is based on the selection of proteins with a particular activity [18, 59–61]. As an example, an exploration of active serine proteases was performed by using chemical probes that interacted with and labeled only active enzymes. The detection of substrates of proteases and kinases and comprehensive analysis of ATP-binding proteins have been reported [18, 59–61]. The last technique has not yet been used for a large-scale proteomics-based search, but possibility of simultaneously monitoring all substrates of a kinase in an intact cell is of thrilling interest for modeling.

The available proteomics tools for studying PTMs, PPIs, and the evaluation of protein activities in many cases fulfill minimal requirements for modeling—for example,

description of a number of components and relation between them. Further standardization of data generation, as well as insights into dynamics of the functional status of proteins, will significantly increase the value of proteomics data.

Localization of Proteins

Cells are spatially organized structures, with intracellular processes being localized to specific organelles and areas of cells. The most informative would be data about protein localization in an intact cell, though it is challenging technically. Currently, protein localization can be evaluated by microscopy techniques of tagged selected proteins in living cells, or by analysis of purified organelles using expression proteomics.

Various microscopy-based techniques have been employed to monitor GFP- or YFP-tagged proteins in yeast and human cells [62–64]. The most comprehensive work has been performed with yeast proteins, with definition of 22 categories for subcellular localizations of 75 % of yeast proteins [62] and the description of the subcellular localization of 6100 proteins [65]. Co-localization of selected proteins can also be monitored in an intact cell by FRET [66]. Co-localization of signaling proteins has been known as an important condition for efficient signaling [67]. Stochastic models of PPI have shown that co-localization is a feature of functional complexes of proteins which do not necessarily form physical complexes [68]. For proteomics studies, it indicates a requirement for the description of protein localization even within a single intracellular compartment—for example, cytoplasm or the nucleus. Thus, a comprehensive definition of protein localization can be achieved by a combination of microscopy-based techniques that describe localization of proteins, with proteomics studies that describe functional features of proteins.

Organelle-specific localization of proteins is important for the modeling of functional relations between protein species. Studies of protein composition of organelles have been among the most successful proteomics works with plant models and are described elsewhere (see Part V in this book). These works provide a solid basis for modeling functions of organelles—for example, chloroplasts and mitochondria.

There is not yet a technique to evaluate protein localization and generate data on the level that is required for systems biology tools. For plant proteomics, microscopic high throughput would have to be developed. Another crucial issue is that well-performed localization studies often lack an evaluation of expression, functional status, and dynamics of proteins. Combinations of various proteomics techniques may solve this issue; at least primary techniques are available—for example, microscopy, 2-DGE, LC–MS/MS, and MS.

Dynamics of Proteomes

The time scale of biological processes involving proteins vary from milliseconds to days. Time scale of a proteomics experiment is therefore defined by a process that is studied. Those processes can be related to changes of only a few proteins or to complex regulatory intracellular processes (e.g., cell differentiation, transformation, and motility), and to developmental processes.

The shortest time scale is characteristic for chemical reactions with proteins; PTMs and PPIs occur in milliseconds to seconds [69–72]. An accumulation of modified proteins may take a longer time, though accumulation rates are dependent on parameters of fast reactions with single proteins. The longer time period, as compared to single-protein modifications, is the result of rearrangements of interactions between proteins and movements of proteins [70, 71]. Signaling processes consist of sequential and parallel changes of various protein species. Every step in such a signaling process requires time for transformation of protein species (e.g., by PTMs), and for rearrangements of the PPIs. The architecture of connections between protein species was found to be crucial for the formation of thresholds, oscillations, and bi-stabilities [8, 9, 12, 72, 73]. Thus, a number of steps in regulatory cascades influence the dynamics of regulatory processes.

Protein movements constitute another factor that defines the dynamic features of proteomes. Substrates and components of complexes have to reach intracellular sites where reactions of modifications and complex formation take place, and they have to be transported to sites of further signal propagation. Estimated diffusion coefficients for most of biologically active proteins are $0.1\text{--}1.0\ \mu\text{m}^2/\text{s}$. It means that a protein may be homogeneously distributed within a cell in less than 10 min [63]. However, cells are not homogeneous chemical reactors. Compartmentalization of cells creates structural barriers for free protein diffusion, and the activities of transport systems affect the duration of signaling processes by introducing time delays, from seconds up to hours [8, 9, 12, 63, 64, 73]. The lack of comprehensive large-scale studies of protein transport systems has been hampering an efficient employment of distribution coefficients in modeling efforts. Thus, the kinetics of reactions with single protein molecules, the number of reactions and connections between protein species, and the movements of proteins constitute three factors that define dynamic properties of proteomes.

Most of the proteomics experiments with time frames longer than 24 h explore the long-term transformations of cells [75, 76]. These effects may be related to cells of a selected type (e.g., exploring the differentiation of cells), or to cells in a context of multicellular organisms (e.g., the study of a developing body). This includes the profiling of various tissues and different cell lines that originate from the same tissue [75, 76]. Thus, longer time frames in proteomics studies are associated with the larger complexity of the studied system. The same paradigm is valid for modeling: Large systems with multiple time-delaying connections require analysis over longer time periods, as compared to systems with a small number of species [8, 70, 77, 78].

Thus, a number and frequency of data collection points can be estimated on the basis of knowledge of studied processes and proteomics technology that will be used. To unveil systemic properties, it may be needed to adjust time collection points to predicted behavior of the system—for example, to collect data at points when significant changes of protein expression would be expected. As an example, fast signaling processes would require data collections within time intervals of minutes over a period of few hours, while for cell differentiation it may be intervals of hours over a period of days.

46.4 REPRESENTATION OF INFORMATION ABOUT PROTEINS AND THEIR INTEGRATION INTO SYSTEMS BIOLOGY TOOLS

If we want to perform a large-scale analysis of proteomics data or would like to design proteomics experiments to study systemic properties of a model, what should be considered? The approach would be to select a proteomics technique that would provide information about required number of protein/species with description of their expression, functional status, and localization (Figure 46.1). The information would have to be collected upon a time period to describe dynamics of changes. Then, collected results would have to be presented in a modeling-compatible format. Finally, modeling tools that would be able to accept levels of details about species and dependencies between them will be selected. In the earlier sections of this chapter, we presented a short overview of requirements of systems biology and what proteomics techniques can deliver. In this section, we will discuss possible ways of integration of systems biology into proteomics studies.

Presentation of Proteomics Data

Tools for acquisition and presentation of biological data in a format that can be used for modeling have been under development. Systems Biology Workbench represents one of such initiatives, with XML-based Systems Biology Markup Language (SBML) being an important step in development of common standards for description of biological data sets—for example, proteomics, genomics, metabolomics, and in-depth studies [79].

The information about a protein has to provide a unique identifier for a protein (e.g., a unique name or tag), and to describe the features of a protein (e.g., molecular mass, PTMs, and its functional status). It also has to describe technical details of how the information was generated, and to be in a format which is compatible with modeling tools. Development of XML-supporting formats for the representation of proteomics data is an important step toward addressing these requirements. MIAPE, PSI-MI, mzXML, AGML, and similar initiatives have laid the groundwork for the systematic recording of proteomics experiments and have been described elsewhere [80–83]. Repository databases that are based on these formats allow documentation of protein identities and technical details of experiments. Unified representation in XML format also simplifies the importation of proteomics data into modeling tools. XML-based SBML is the basis for representation of biological models, and it allows exchange of data between different models in a well defined format. The unified representation requires adoption of a nomenclature for protein species and their presentation in functional states—for example, PTMs status and interactions with other proteins. The GO project provides nomenclature [84] that can be the basis for nomenclature of protein species, because GO-adopted protein names can be core names for the description of modified proteins and proteins in complexes (e.g., protein species). Thus, when planning a proteomics experiment, standardized annotation of proteins and their deposition in modeling-compatible formats have to be considered.

Architecture of a Model and Description of Dependencies

Repository databases provide a framework for the description of static features of proteins, while tools for a systematic description of protein dynamics are less developed. The latter requires connecting protein species with defined PTMs in defined complexes and in a defined location, through a given time interval. This task can be solved in a framework of modeling formats, similar to formats supporting SBML. These formats incorporate description of the functional, spatial, and quantitative features of proteins (see www.sbml.org for examples) [79]. It is important to note that the SBML and related efforts allow the building up and expansion of an existing model upon generation of new data.

Modeling tools that can accommodate data of various degree of definition of details about components have been described [8, 9, 12]. High-degree definition models are based on a precise knowledge of (a) features of data sets that are sufficiently rich to build a system of defined dependencies between variables and (b) precisely calculated values of variables in dynamics. In mathematical terms, such models are based on differential equations. Because the use of differential equations requires detailed input, such modeling has been successfully applied to functional modules, but not to cell-wide data sets. Common to such models is the requirement for information about concentrations and kinetic parameters of all involved protein species. Lack of such information may lead to collapse of the model or to incorrect conclusions.

To overcome the requirement of high definition, modeling tools based on correlative relations have been developed. For instance, Boolean networks can be built with data about connections and the functional status of species of a system. They also consider the number of inputs and outputs for components. Boolean networks can analyze data that describe the state of a single variable and the average state of all other neighbors [85–87]. They do not require knowledge about concentrations and the kinetic parameters of components. Bayesian networks allow analysis of even less well-defined data, because they describe statistical relationships between species from a data set, and do not require knowledge of confirmed connections between species. These statistical relationships may include pairwise relations and relations between arbitrary complexes. Bayesian networks summarize dependencies between variables and select for the highest probable model, given the data [88]. Essentially, Boolean and Bayesian networks are based on the statistical values of relations that may be exemplified by the statement “If A and B are up, then C should be down”. For proteomics data sets, it provides an opportunity to search for proteins with correlative behavior—for example, expression, pattern of phosphorylation, and so on.

Statistical analysis of various types has also been applied to data sets with a different definition of details about components, increasing a number of modeling approaches [89, 90]. Thus, the type of proteomics data and biological questions to be answered would define which modeling tools to be employed; the number of proteins/species, their functional status and dynamics, and the complexity of a studied process would define the architecture of the model, define the mathematical tools to describe dependencies between species, and predict behavior of the model.

Input from Modeling of Abstract Systems

Modeling of abstract systems suggests the most optimal size, relations, and dynamics of species that are conditions for acquiring of such systemic features as robustness and sensitivity. Abstract mathematical modeling defines restriction rules for the modeling of biological systems: For instance, it describes how stability of a system will be affected by a number of affected species and by a number of inputs and outputs of these species [86, 87, 91, 92]. Moreover, it also predicts the requirement of a certain level of noise, as random and nonspecific changes in dependencies and values, in order to have a system evolving [86]. Knowledge of the behavior of abstract mathematic models sets critical limits for quantity and features of species and improves the computing of biological models.

46.5 CONCLUSIONS

Proteomics has reached the level when generated results are difficult to evaluate without a large-scale analysis. Proteomics data sets contain information about hundreds of affected species and about their functional status and dynamics, and all observed changes can be correlated with cellular processes. Such rich data sets require unbiased analysis that would evaluate all species and not simply concentrate on only one or two selected proteins. Tools for such an analysis are available in systems biology. These tools allow processing of high- and low-definition data and provide possibilities to model cell responsiveness. Developments in proteomics and systems biology bring these two fields together and can produce more novel information than each of the fields would have delivered alone.

46.6 FIVE-YEAR VIEWPOINT

Proteomics and systems biology are two fast growing fields. In five years, application of systemic analysis to proteomics data will be a routine procedure. This will be possible due to following developments: (i) availability of tools for an optimal design of an experiment predicted to unveil systemic properties of a biological model; (ii) further development of proteomics for comprehensive and detailed description of proteomes in their dynamics; (iii) standardized acquisition and presentation of proteomics data; (iv) primary analysis of data using comprehensive databases; and (v) automated selection of modeling tools corresponding to the type of proteomics data.

ACKNOWLEDGMENTS

I apologize for not being able to cite all relevant works due to limitations of space. Works in the author's laboratory are supported in part by grants from the Karolinska Biomics Centre, Ludwig Institute for Cancer Research, the Swedish Cancer Society, the Swedish Research Council, Hiroshima University, INTAS, EU RTN EpiPlast Carcinoma, and Merck KGaA.

REFERENCES

1. Purves WK and Orians GH (eds.). 1987. *Life: The Science of Biology*, 2nd edition, Sinauer, Sunderland MA.
2. Cavnas FM, et al. 2004. *Proteomics* **4**:285–298.
3. Chen S and Harmon AC. 2006. *Proteomics* **6**:5504–5516.
4. de Hoog C and Mann M. 2004. *Ann Rev Gen Hum Genet* **5**:267–293.
5. Görg A, Weiss W. and Dunn MJ. 2004. *Proteomics* **4**:3665–3685.
6. Souchelnytskyi S. 2005. *Proteomics* **5**:4123–4137.
7. Glinski M and Weckwerth W. 2006. *Mass Spectrom Rev* **25**:173–214.
8. Eungdamrong NJ and Iyengar R. 2004. *Biol Cell* **96**:355–362.
9. Cho K-H and Wolkenhauer O. 2003. *Biochem Soc Trans* **31**:1503–1509.
10. Kitano H. 2004. *Nature Rev* **4**:227–235.
11. Bhaskaran N and Souchelnytskyi S. 2008. *Proteomics* **10**:In press.
12. Ideker T and Lauffenburger D. 2003. *Trends Biotechnol* **21**:255–262.
13. Schoeberl B, et al. 2002. *Nat Biotechnol* **20**:370–375.
14. Kutalik Z, Cho K-H, and Wolkenhauer O. 2004. *BioSystems* **75**:43–45.
15. Blagoev B, et al. 2004. *Nat Biotechnol* **22**:1139–1145.
16. Dubrovskaya A and Souchelnytskyi S. 2005. *Proteomics* **5**:4678–4683.
17. Cutler P. 2003. *Proteomics* **3**:3–18.
18. Adam GC, Sorensen EJ, and Cravatt BF. 2002. *Mol Cell Proteomics* **1**:781–790.
19. Vosseller K, et al. 2005. *Proteomics* **5**:388–398.
20. Zhou H, Watts JD, and Aebersold R. 2001. *Nat Biotechnol* **19**:375–378.
21. Oda Y, Nagasu T, and Chait BT. 2001. *Nat Biotechnol* **19**:379–382.
22. Tonge R, et al. 2001. *Proteomics* **1**:377–396.
23. Martin K, et al. 2003. *Proteomics* **3**:1244–1255.
24. Steinberg TH, et al. 2001. *Proteomics* **1**:841–855.
25. Gygi SP, et al. 1999. *Nat Biotechnol* **17**:994–999.
26. Ong SE, Foster LJ, and Mann M. 2003. *Methods* **29**:124–130.
27. Julka S and Regnier F. 2004. *J Proteome Res* **3**:350–363.
28. Schmidt A, Kellermann J, and Lottspeich F. 2005. *Proteomics* **5**:4–15.
29. Gevaert K, et al. 2003. *Nat Biotechnol* **21**:566–569.
30. Reinders J, et al. 2004. *Proteomics* **4**:3686–3703.
31. Traxler E, et al. 2004. *Proteomics* **4**:1314–1323.
32. Gerner C, et al. 2002. *Mol Cell Proteomics* **1**:528–537.
33. Greighton TE. 1993. *Proteins: Structure and Molecular Properties*, W. H. Freeman and Company, New York, pp. 78–102.
34. <http://www.abrf.org>
35. <http://www.expasy.org>
36. Pawson T. 2004. *Cell* **116**:191–203.
37. Heldin C-H and Purton M. 1996. *Signal Transduction*, Chapman and Hall, London, pp. 1–365.
38. Guy GR, Philip R and Tan YH. 1994. *Electrophoresis* **15**:417–440.

39. Stasyk T, et al. 2005. *Mol Biol Cell* **16**:4765–4780.
40. Knight ZA, et al. 2003. *Nat Biotechnol* **21**:1047–1054.
41. Hirabayashi J and Kasai K. 2002. *J Chromatogr B Anal Technol Biomed Life Sci.* **771**:67–87.
42. Peng J, et al. 2004. *Nat Biotechnol* **21**:921–926.
43. Vertegaal ACO, et al. 2004. *J Biol Chem* **279**:33791–33798.
44. Fu M, et al. 2002. *Cytokine Growth Factor Rev* **13**:259–276.
45. Umlauf D, Goto Y, and Feil R. 2004. *Methods Mol Biol* **287**:99–120.
46. Ge Y, et al. 2004. *Proteomics* **4**:3464–3467.
47. Bork P, et al. 2004. *Curr Opin Struct Biol* **14**:292–299.
48. Gavin A-C, et al. 2002. *Nature* **415**:141–147.
49. Ho Y, et al. 2002. *Nature* **415**:180–183.
50. Bouwmeester T, et al. 2004. *Nat Cell Biol* **6**:97–105.
51. Giot L, et al. 2003. *Science* **302**:1727–1736.
52. Li S, et al. 2003. *Science* **303**:540–543.
53. Colland F, et al. 2004. *Genome Res* **14**:1324–1332.
54. von Mering C, et al. 2002. *Nature* **417**:399–403.
55. Gagneur J, et al. 2004. *Genome Biol* **5**:R57.
56. Park OK. 2004. *J Biochem Mol Biol* **37**:133–138.
57. Pan S, Carter CJ and Raikhe NV. 2005. *Exp Rev Proteomics* **2**:781–792.
58. Rhee SY, Dickerson J, and Xu D. 2006. *Annu Rev Plant Biol* **57**:335–360.
59. Saghatelian A, et al. 2004. *Proc Natl Acad Sci USA* **101**:10000–10005.
60. Bredmeyer AJ, et al. 2004. *Proc Natl Acad Sci USA* **101**:11785–11790.
61. Besant PG, et al. 2004. *J Proteome Res* **3**:120–125.
62. Huh WK, et al. 2003. *Nature* **425**:686–691.
63. Roix J and Mistelli T. 2002. *Histochem Cell Biol* **118**:105–116.
64. Liebel U, et al. 2003. *FEBS Lett* **554**:394–398.
65. Kumar A, et al. 2002. *Genes Dev* **16**:707–719.
66. Mahajan NP, et al. 1998. *Nat Biotechnol* **16**:547–552.
67. Hlavacek WS, et al. 2006. *Sci STKE* **344**, online issue.
68. Batada NN, Shepp LA, and Siegmund DO. 2004. *Proc Natl Acad Sci USA* **101**:6445–6449.
69. Noble D. 2002. *BioEssays* **24**:1155–1163.
70. Schoeberl B, et al. 2002. *Nat Biotechnol* **20**:370–375.
71. Ballif BA, et al. 2005. *Proc Natl Acad Sci USA* **102**:667–672.
72. Ishii N, et al. 2004. *Biotechnology* **113**:281–294.
73. Hasty J, McMillen D, and Collins JJ. 2002. *Nature* **420**:224–230.
74. Vo TD, Greenberg HJ, and Palsson BO. 2004. *J Biol Chem* **279**:39532–39540.
75. Behre G, et al. 2002. *Expert Opin Ther Targets* **6**:491–495.
76. Souchelnyskyi S. 2002. *J Mammary Gland Biol Neoplasia* **7**:339–371.
77. Ideker T, et al. 2001. *Science* **292**:929–933.
78. Yan W, et al. 2004. *Genome Biol* **5**:R54.
79. Hucka M, et al. 2003. *Bioinformatics* **19**:524–531.
80. Hermjakob H, et al. 2004. *Nat Biotechnol* **22**:177–183.

81. Taylor CF, et al. 2003. *Nat Biotechnol* **21**:247–254.
82. Pedrioli PGA, et al. 2004. *Nat Biotechnol* **22**:1459–1466.
83. Stanislaus R, et al. 2004. *BMC Bioinformatics* **5**:1–7.
84. <http://www.geneontology.org>
85. Amaral LAN, et al. 2004. *Proc Natl Acad Sci USA* **101**:15551–15555.
86. Kauffman S, et al. 2004. *Proc Natl Acad Sci USA* **101**:17102–17107.
87. Huang S and Ingber DE. 2000. *Exp Cell Res* **261**:91–103.
88. Sachs K, et al. 2002. Science's STKE <http://stke.sciencemag.org/cgi/content/full/sigtrans;2002/148/pe38>
89. Plavec I, et al. 2004. *Proc Natl Acad Sci USA* **101**:1223–1228.
90. Spirin V and Mirny LA. 2003. *Proc Natl Acad Sci USA* **100**:12123–12128.
91. Angeli D, Ferrell JE, and Sontag ED. 2004. *Proc Natl Acad Sci USA* **101**:1822–1827.
92. Song C, Havlin S, and Makse HA. 2005. *Nature* **433**:392–395.

PROTEOMICS IN DEVELOPING COUNTRIES

Nat N. V. Kav, Sanjeeva Srivastava, William Yajima,
and Shakir Ali

47.1 INTRODUCTION

Proteomics

Genomics, which is the study of the genetic content of a particular cell or organism, has revolutionized scientific research. The success of various genome sequencing projects, including the complete annotation of the genome sequences of plant species such as *A. thaliana* and rice as well as large-scale EST sequencing initiatives <http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html>, have provided the scientific community with a wealth of information in the areas of plant physiology, biochemistry, and morphology. The next logical endeavor for the scientific community involves the structural and functional characterization of the products of the identified genes. Those products are proteins, and the burgeoning area of scientific research involving the elucidation of their various roles is known as proteomics. The term “proteome” was first used in the mid-1990s to describe the total set of proteins expressed in a given cell at a given time, and the word was derived from the fact that the study of proteomics is basically the study of the protein complement of the genome [1].

Proteomics can essentially describe any research that involves (a) the detailed characterization of proteins, and (b) the characterization of the proteome, including levels of expression, PTMs, localization, and interactions with other molecules in order to obtain a global view of cellular processes at the proteome level [2]. Research in this field of study is directed toward analyzing protein dynamics within an organism, tissue, or cell and is becoming quite common in plant biology to decipher the functions of various genes. Proteomics can therefore be for the discovery of novel targets (discovery proteomics) or the characterization of the structure and function of proteins (functional proteomics).

One of the most commonly utilized techniques in “discovery proteomics” to profile the proteomes of cells and organelles consists of protein extraction and separation by 2-DGE or by LC [3–5]. 2-DGE is a PAG-based protocol that separates proteins on the basis of two distinct parameters: molecular weights, and pI [6]. It is possible for this procedure to efficiently separate more than 1000 proteins stained with silver in an individual PAG [7]. Although 2-DGE is one of the classical and most widely used methods of proteome analysis [8], there are certain inherent disadvantages to this method such as low sensitivity, difficulty in resolving proteins that are large (over 150 kDa) or small (less than 10 kDa), or proteins that have very low (less than 3) or very high (greater than 10) pI values and the inability to resolve proteins with low solubility [3, 9, 10]. LC is a column-based procedure that uses chromatographic resins to separate peptides and proteins based on differences in specific characteristics such as affinity for specific ligands, size, ionic character, or hydrophobicity. It can be used as a prefractionation technique to reduce the complexity of a protein sample or to study the protein content of specific cellular components such as membranes or organelles. LC or 2D-LC has been demonstrated to be very reproducible, and crude protein extracts can be analyzed after only a few purification steps [11]. LC-based methods have been used to characterize large numbers of plant proteins [12, 13]. Selected protein spots from 2D gels or LC-protein fractions can be digested with enzymes that cleave at specific peptide bonds and then subjected to MS analysis for qualitative and quantitative comparisons. DIGE is another proteomics research technique and is quite similar to 2-DGE; however, there are differences because this type of gel electrophoresis involves tagging two protein solutions to be analyzed with different fluorescent dyes, such as Cy2, Cy3, or Cy5, and then mixing the solutions and analyzing them on the same gel [14, 15]. This allows for the visualization of any differences in protein amounts between the analyzed samples based on the predominance of one fluorescent dye over the other. In addition to 2-DGE and LC, there are other techniques that can be used in proteomics studies including MudPIT (which has been shown to be especially effective when attempting to identify as many proteins as possible from a particular sample [16]), ICAT, iTRAQ for relative and absolute quantitation, and SILAC [9, 17–21]. Such discovery techniques are discussed in detail elsewhere in this book and are beyond the scope of this chapter.

Discovery of Novel Targets vs “Functional” Proteomics

While there is considerable interest and research focused on EST or transcriptomics analyses (microarrays, gene chips), which involve investigations at the level of

transcripts, there are certain limitations to these studies. In particular, there is not always a correlation between changes in the levels of mRNA and protein quantity within a cell [22, 23]. Therefore, in addition to transcriptional studies, research into the structure and function of the encoded proteins is required in order to obtain comprehensive and accurate information on cellular activities. A major focus of plant proteomics research has been discovery-proteomics or abundance-based proteomics, which involves the identification of proteins involved in mediating various plant processes such as abiotic stress tolerance or disease resistance. The discovery-based proteomics approaches have proved their utility as demonstrated by the identification of novel targets using a variety of approaches [24]; however, the mere identification and subsequent validation of proteins are not sufficient to understand the function of each of the identified proteins. In order to fully comprehend such complex machineries, a thorough understanding of protein structure and function as well as how various proteins interact and function in concert is required. Therefore, another focus of proteomics research involves the study of functional proteomics, which attempts to generate detailed biological information on the role(s) played by each protein identified in the discovery phase. Discovery proteomics and functional proteomics have been described as forward proteomics and reverse proteomics, respectively [3]. Forward proteomics might be considered the classical or traditional form of proteomics research, and it consists of the identification and characterization of proteins from a particular sample [3]. This approach to proteomics research involves the familiar sequence of procedural techniques of protein extraction, fractionation, or separation and subsequent mass spectrometry analysis for protein identification. Reverse proteomics requires the isolation of specific cDNA sequences for proteins of interest, protein overexpression, and subsequent experimental assays to determine functionality of the corresponding protein [3]. Functional proteomics has emerged as a powerful tool to investigate the interaction, integration, and function(s) of proteins in specific plant processes by providing information on protein PTMs or PPIs [25, 26]. The use of Y2H experiments has led to significant advances in the understanding of PPIs in yeast [27], *C. elegans* [28], *D. melanogaster* [29, 30], and humans [31].

A technology that has the potential for high-throughput identification of proteins utilizes protein microarrays, which can allow researchers to study protein expression profiles and monitor global sets of proteins produced by any given cell type, tissue, or organism [32, 33]. Protein microarrays consist of capture agents including proteins, antibodies, or peptides that selectively bind target proteins, immobilized on a solid surface providing a high-throughput platform for the characterization of PPIs. The ideal functional protein array would allow the concomitant study of all of the proteins encoded by the genome of an organism. Such an array would be the whole proteome equivalent of the whole genome arrays that are currently being used for investigations at the transcript level. Arrays containing families of proteins with similar biological activities (e.g., protein kinases), proteins expressed in the same tissue, or proteins associated with a particular response (e.g., incompatible interactions with pathogens) may also have obvious utility for plant scientists. In the future, technical advances in protein chip and MS techniques combined with more sophisticated and comprehensive protein databases will greatly assist in the detailed study of a plant proteome [34].

A more detailed description of proteomics-based techniques is reviewed in Part I: Technologies.

Proteomics in Crop Improvement

The Need for Crop Improvement. Optimization of crop yield and quality are perhaps the two most important factors for growers; therefore, crop improvement involving the engineering of plants is an area that will likely continue to be the focus of a significant amount of research as well as funding. In addition to the rapidly increasing global population, several factors including social, geopolitical, economical, and environmental factors have raised concerns regarding the ability to produce a sufficient amount of high-quality, nutritious food in an efficient and sustainable manner throughout the world. It was estimated in the late 1990s that about two billion people worldwide suffer from malnutrition and/or hunger; this figure is likely to increase as the global population continues to grow at an alarming rate, with a projected world population of over nine billion by 2050 and with much of the increase occurring in developing nations [35]. Although it is estimated that approximately 50–80% of the population in developing countries are involved in some capacity in the agricultural and farming sector, these nations have the lowest agricultural output in terms of crop yield; not unexpectedly, much of their population subsists on a diet that is nutritionally inadequate [35]. Furthermore, various abiotic and biotic stresses are affecting plant growth and food productivity worldwide. Abiotic stresses such as salinity and drought affect crop production, and a loss of 30% of arable land by 2025 and 50% by 2050 due to increased salinity stress is expected [36]. The earth has 25% arid and desert land (dry land salinity), and 20% of the irrigated land is also affected by increased salinization (irrigated land salinity) [37]. It is estimated that more than 800 million hectares of land throughout the world is currently affected by salinity [38]. Abiotic stresses are, in fact, a principal cause of crop failure worldwide, and it threatens the sustainability of agricultural outputs by limiting crop productivity. Crop losses due to biotic factors are no less significant as demonstrated by the fact that in the mid-1990s, it was estimated that about \$500 billion (US) worth of crops were lost each year globally due to weeds, pests, and diseases and that losses were significantly greater in developing nations compared to developed countries [35]. Annual worldwide losses in crops such as barley, coffee, cotton, maize, potato, rice, soybean, and wheat due to insects, weeds, and disease can reach approximately 40% [35]. In addition, there are increasing demands on the available, arable land to produce non-food crops such as biofuels, other bioproducts, and therapeutic agents. Producers may respond by choosing to cultivate these non-food crops for the production of biodiesel, bioplastics, biomass, and other bioproducts because higher profit margins may be associated with such commodities. Therefore, abiotic and biotic factors that reduce crop yield and increase demands on available land for the cultivation of plants for non-food uses pose major challenges to meet the predicted increase in demand for safe, nutritional food for a growing global population. Clearly, increasing the yield of food crops in normal, arable land available for food production as well as increasing the range of crop plants to produce high yields on less productive lands (e.g., saline soils) is of

paramount importance. Proteomics, therefore, may provide important information at the molecular level about crucial processes that mediate plant responses to abiotic and biotic stresses which can be used for rational crop improvement.

Conventional Breeding vs Genetic Engineering. The ability of a plant to tolerate abiotic and biotic stresses or to be high yielding—or, for that matter, to have any other desired characteristic—may be improved through the manipulation of a plant's genetic content. Such improvements to crop varieties may be made through classical breeding techniques or by genetic engineering. Classical plant breeding takes advantage of homologous recombination of genetic material that can occur during naturally occurring mating processes, and it involves the crossing of related plant species in order to create new varieties that possess desirable traits [39]. For example, a crop variety that exhibits an ability to tolerate salinity might be crossed with a related high-yielding variety in order to develop new high-yielding, saline-tolerant varieties. Because of the nature of conventional plant breeding, the breeder does not know precisely which genes from either parent are being introduced into newly developed varieties. However, genetic modification of crops using genetic engineering techniques usually involves the introduction of a handful of well-characterized genes [40, 41]. Thus genetic engineering of crops for the production of new varieties with enhanced traits can proceed much faster and with a greater degree of accuracy compared to conventional breeding [42]. Although conventional plant breeding utilizes naturally occurring mating processes, *in vitro* experimental procedures including embryo rescue, protoplast fusions, and mutagenesis with radiation and chemicals such as EMS and DMSO are essential tools often used by conventional plant breeders [39]. The information garnered from genomics and proteomics-based studies has furthered our understanding of the genetic and biochemical basis of plant abiotic and biotic stress tolerance, which in turn has accelerated the generation of transgenic plants exhibiting tolerance to these stresses [43]. It is expected that plant proteomics (and genomics) will continue to have a major role in crop genetic engineering by identifying additional targets that can be manipulated. Some of our own studies described below provide evidence to this statement.

From Target Discovery to Validating Its Utility: Our Experience. The Kav laboratory at the University of Alberta, Canada routinely uses proteomics and genomics to investigate plant responses to abiotic and biotic stresses. For example, Kav et al. [44] identified root proteins affected by salinity in pea (*P. sativum*) from control and salinity-treated plants and seedlings by using 2-DGE and MS. The identities of these proteins included PR10 proteins and antioxidant enzymes such as SOD and NDPK [44], and this investigation suggested a potential role for PR10 proteins in salinity stress responses. In addition, the proteome of a saline-tolerant *A. hypogaea* L. callus cell line was compared with its sensitive counterpart, which revealed many PR10 proteins as being unique or significantly up-regulated in the tolerant line [45]. Based on these proteomics-based findings, it was hypothesized that PR10 proteins play a crucial role in mediating plant responses to salinity stress. In order to test this hypothesis, transgenic *B. napus* as well as *A. thaliana* lines that constitutively expressed two

different members of the pea PR10 family—PR10.1 and ABR17 respectively—were generated. The transgenic *B. napus* [46] and *A. thaliana* [47] plants exhibited enhanced germination, early flowering, and tolerance to multiple abiotic stresses, supporting the hypothesis that PR 10 proteins play crucial roles in stress tolerance. We have also used proteomics to investigate differences between auxinic herbicide-susceptible and -resistant biotypes of wild mustard (*S. arvensis* L.), which led to the identification of a peptidylprolyl *cis*–*trans* isomerase [48]. Other examples of application of proteomics to identify novel targets in the Kav laboratory included the identification of several proteins that may have important roles in the resistance of *Brassica* species to the pathogens *L. maculans* [49] and *A. brassicae* [50]. The Kav laboratory has also applied proteomics to characterize the ubiquitous plant pathogen *S. sclerotiorum*, capable of infecting more than 400 different plant species; this study led to the identification of a potential novel virulence factor that had not been identified by transcript analysis [51]. Therefore, it is evident from the examples from our laboratory that proteomics has a very important role in identifying proteins (hence genes) that play crucial roles in mediating plant responses to stress, and the manipulation of those genes may lead to novel crop varieties with enhanced tolerance to stress.

Proteomics for Investigating Transgene-Mediated Changes in Plant Proteomes. Using information garnered from proteome-level studies to create new transgenic plants that are better able to tolerate adverse environmental conditions or disease-inducing biotic factors must be performed with the understanding that the harvested crop must ultimately be safe for consumption. Genetic manipulations have the potential to cause unintended effects such as increased toxicity and allergenicity; therefore, the concept of substantial equivalence was developed as one of the parameters to be used to evaluate whether a new transgenic crop is likely to be safe for ingestion. If a new transgenic crop is considered to share the same nutritional and health characteristics with its most closely related, naturally occurring crop or food, which has been used safely as a food product, then the newly developed transgenic crop is also considered to be safe for human consumption [52]. The determination of substantial equivalence can be based on how similar the metabolic and proteomic profiles are between the novel crop and the existing crop. Techniques such as PCR, quantitative real-time PCR, and HPLC analysis have been used for the comparative analyses of intended effects caused by genetic transformation(s) [53], whereas approaches using cDNA microarrays and proteomics techniques can be used to identify even unintended effects.

The substantial equivalence of a genetically modified variety of potato and a transgenic tomato plant that displayed increased viral resistance was determined, and in both cases the novel transgenic plants were found to meet the criteria of substantial equivalence and were subsequently deemed to not pose any greater risk than their naturally occurring wild-type varieties [54, 55]. Similarly, to investigate the unintended insertional and pleiotropic effects, 12 transgenic *A. thaliana* lines were analyzed by 2-DGE and the seed proteomes of these lines were compared with their parental lines. Results from these analyses demonstrated that unintended changes did not take place in the transgenic lines [56]. Microarrays are another important tool for the assessment of substantial equivalence. Six independently transformed *A. thaliana*

lines modified in the flavonoid biosynthesis pathway were tested by cDNA microarrays to detect any adverse and unintended consequences of the genetic alteration [53]. Microarrays have also been used to compare the expression in nontransgenic and transgenic wheat expressing a fungal phytase gene [57] and a gene encoding HMW subunits of glutenin [58]. Results from these studies also suggested no significant difference in gene expression between nontransgenic and transgenic lines. In fact, differences between conventionally bred materials were much larger as compared to differences between nontransgenic and transgenic lines [58]. Thus techniques capable of characterizing global changes in gene expression including proteomics and microarrays may be extremely useful in identifying unintended, potentially deleterious effects.

47.2 PROTEOME RESEARCH IN THE DEVELOPING WORLD

Plant Proteomics Research in Developing Countries

Although the general perception of countries that are considered developing as opposed to developed may be that they do not possess the facilities, technology, and money necessary for performing sophisticated proteomics-based research, it is clear that this is not the case because there are a number of examples of significant contributions being made by scientists in developing countries to the field of plant proteomics. Proteome-level research conducted at the Agricultural Biotechnology Research Institute of Iran in collaboration with universities such as Ferdowsi University of Mashhad and Tehran University have allowed for a greater understanding of how sugar beets [59], *S. aegyptiaca* [60], rice [61], and *E. elongatum* [62] are affected by various abiotic stresses. Thai researchers at the King Mongkut's University of Technology have also studied the proteome-level differences among bran proteins of 14 varieties of rice [63] while a bioinformatics network (LACBioNet), which has applications in proteomics research, has been developed for Latin America and the Caribbean nations [64].

In addition to India, which will be discussed in much greater detail in subsequent sections, it can be argued that one of the world's biggest developing countries in terms of geographical size and population that possesses the capabilities of performing valuable plant proteomics-based research is China. It is believed that proteomics-based studies in China evolved from earlier classical protein biochemistry studies and began in earnest in the late 1990s with the initiation of proteomics-related projects at the Hunan Normal University, the Shanghai Institutes of Biological Sciences of Chinese Academy of Sciences, the National Centre of Biomedical Analysis of China, and the Beijing Institute of Radiation Medicine, all with the support of the National Natural Science Foundation of China [65]. Since its inception, proteomics-based research in China has expanded to a number of other facilities, including the Institutes of Biomedical Sciences at Fudan University, the College of Life Science at Peking University, and the Institute of Basic Medicine at the Chinese Academy of Medical Sciences [65]. In October 2005, various levels of Chinese governments along with a number of leading research facilities such as the National Centre for Biomedical Analysis, the

Chinese Academy of Sciences, Peking University, Tsinghua University, the Chinese Academy of Medical Sciences, and Beijing Pharma and Biotech Centre joined to form the Beijing Proteomics Research Centre centralized in the Beijing Zhongguancun Life Science Park, which is the largest single proteomics research station in China [65]. With the support of numerous government funding programs such as the National Core Facilities of Protein Science and the Long-Term Science and Technology Program, it is anticipated that in the coming years proteomics-based studies will continue to flourish in China [65].

Although much of the proteomics-based research that has taken place in Chinese laboratories has involved the study of the human proteome, there has also been considerable attention given to various plant proteomics studies. For example, numerous studies have been performed on rice, including those involving abiotic stresses [66–68], biotic stress [69], and rice growth and development [70–73]. Proteomics-based studies of the physiology of *A. thaliana* [74, 75] and wheat plant host responses to fungal infection [76] have also been completed at various Chinese laboratories. These studies as well as a number of other related studies are more thoroughly summarized in other chapters and are mentioned here only to highlight some of the significant contributions made by researchers in China in plant proteomics.

The following sections of this chapter examine the role of another developing nation, India, in plant-based proteomics research.

Proteome Research and Available Facilities in India

India has had a strong presence in protein research dating back to the early 20th century. For example, Professor G. N. Ramachandran's research work at the Indian Institute of Science (IISc), Bangalore, in the field of crystal physics and crystal optics made significant contributions in the fields of X-ray crystallography, Fourier syntheses, and X-ray intensity statistics [77]. In 1955, Professor Ramachandran and colleagues, working at the University of Madras, discovered the triple helical structure of collagen, the most abundant protein in connective tissue [78, 79]. Furthermore, the "Ramachandran plot" is recognized as one of the most outstanding contributions in structural biology and is widely used in biochemistry textbooks [80]. Following in Professor Ramachandran's tradition, many Indian scientists have continued structural and functional analysis of proteins (functional proteomics) and peptides in various Indian laboratories and institutes [81–89].

As described in the previous paragraph, India has made (and continues to make) significant contributions to the area of "functional proteomics," and facilities to conduct research in this area are available throughout India. However, when it comes to discovery proteomics, India's infrastructure and contributions are perhaps not that well-developed. A summary of known institutes and laboratories as well as their facilities for conducting discovery proteomics research is now provided.

Premier institutes in India such as the Center for Cellular and Molecular Biology, Hyderabad (CCMB) <http://www.ccmb.res.in/>, Institute of Genomics and Integrative Biology, New Delhi (IGIB) <http://www.igib.res.in/>, Central Drug Research Institute, Lucknow (CDRI) <http://www.cdriindia.org/home.asp>, which has a robotic-assisted

high-throughput screening platform (Beckman Coulter) and proteomics platform (Perkin–Elmer), and the All India Institute of Medical Sciences, New Delhi (AIIMS) <http://www.aiims.edu/> have started developing facilities for proteome-level research, and these major research institutions have developed collaborations with other smaller research facilities and laboratories. In addition to the aforementioned research facilities, university departments such as the Department of Biochemical Engineering and Biotechnology at the Indian Institute of Technology (IIT), New Delhi <http://www.iitd.ernet.in/deptt/beb/>, have also initiated proteomics-based research. Pioneering work on protein structure and function in India is being performed at AIIMS using 2-DGE and MS. IGIB is a laboratory associated with the Council of Scientific and Industrial Research (CSIR) <http://www.csir.res.in/>, and it is equipped with a 4-tetra FLOPS Hewlett-Packard Supercomputer running the Linux operating system, which is being used to advance its life science computational biology research. The scientists at IGIB work closely with the scientists at The Center of Genomics Application (TCGA) <http://www.tcga-research.org/>, and this collaboration is helping to facilitate a closer partnership between the public and private sectors for the development of genomics- and proteomics-based research. The facilities at TCGA devoted to proteomics research utilize a 2-DGE system, MALDI–TOF/TOF (Bruker Ultra flex), and nLC–MS (Agilent). The experience and expertise of researchers in India in MS and discovery proteomics to study biomarker and signaling molecules are being effectively utilized in research institutes such as Kasturba Medical College, Manipal <http://www.manipal.edu/kmc/>. At the Industrial Toxicology Research Centre, Lucknow (ITRC) <http://www.itrcindia.org/>, serum and tissue proteome profiling in toxicant-exposed humans and animals is being performed, while other world class laboratories and institutions in India with leading research projects include the Saha Institute of Nuclear Physics, Calcutta (SINP) <http://www.saha.ac.in/cs/www/>, Centre for DNA Fingerprinting and Diagnostics, Hyderabad (CDFD) <http://www.cdfd.org.in/>, IISc, Bangalore (<http://www.iisc.ernet.in/>), National Institute of Immunology, Delhi (NII) <http://www.nii.res.in/>, University of Delhi <http://www.du.ac.in/>, National Centre for Biological Sciences, Bangalore (NCBS) www.ncbs.res.in/, and the Tata Institute of Fundamental Research, Mumbai (TIFR) www.tifr.res.in/.

The government of India has also taken steps to support proteome research initiatives through its various agencies such as the Department of Biotechnology (DBT) <http://www.du.ac.in/> and CSIR. The DBT has set up facilities that have the capabilities of performing MALDI–TOF–MS as well as also setting up clinical proteomics facilities at AIIMS and the National Center for Plant and Genomic Research, New Delhi (NCPGR) <http://www.ncpgr.nic.in/>. Furthermore, a program on “clinical proteomics” whose purpose is to help create a collaborative network of basic researchers and clinicians involved in the application of proteomics technologies from “bench side” to “bedside” has been initiated by the DBT. In India, most of the efforts in proteome-level research are currently focused on medical and clinical studies; however, proteomics-based research programs involving the study of plants and agriculture are gaining momentum.

Bioinformatics and a Leadership Role for India

In recent years, India has utilized its strength in information technology and bioinformatics to contribute to various genome annotation projects. For example, even though India's involvement in genome sequencing projects has been very limited, Indian scientists have participated in the sequencing of rice as well as *E. histolytica* genomes [90, 91]. Therefore it is obvious that the knowledge base necessary for proteome research as well as in the analysis of the data generated using information technology exists in India, and this country can play an important role in plant proteome research in subsequent years.

India's biotechnology revolution (evident when one visits metropolitan centers in India) in the areas of pharmaceuticals, agriculture, clinical research, or bioinformatics is well known [92]. India is an emerging biotechnology leader in the Asia-Pacific region. According to the Ernst and Young Global Biotech Report, India is expected to generate \$5 billion in revenues and create more than one million biotechnology jobs over the next five years. Rajiv Memani, CEO and Country Managing Partner for Ernst and Young India, states, "Analysis of biotech events in India clearly indicates that Indian biotech companies are getting their fundamentals firmly in place, business models are maturing, and product commercialization capabilities are improving" [93]. To further substantiate his claims, one can examine what is occurring in the city of Bangalore, which, in India, is one of the top emerging biotechnology centres. The 100-acre biotechnology park near the Electronic City in Bangalore is currently under construction and was expected to be operational by the end of 2006 [94]. The growth of the biotechnology sector in India can be demonstrated by the fact that it is estimated that two new biotechnology companies are being set up every month in the Bangalore biotechnology park while Karnataka has attracted the highest venture capital fund in Asia (~\$100 million) [95]. Kiran Mazumdar Shaw, who is the chairperson of Vision Group on biotechnology and CMD, Biocon Ltd, believes that "Bangalore is evolving as an interesting ecosystem for biotechnology" [96]. Other states in India are using what is occurring in Bangalore as a model in the hopes of establishing their own niche in the rapidly expanding biotechnology field. "Bangalore Bio" has emerged as a key event in India to advertise and inform those who are interested in what is occurring in the biotechnology industry in India. The attendance of national and international speakers, hundreds of conference delegates and exhibitors, and many thousands of visitors from various countries at this event serves notice that the recent biotechnology boom in India is viable and will likely continue to expand in the near future. The Bangalore Bio event also provides Indian researchers an opportunity to network with international scientists in the hopes that fruitful collaborations might develop with researchers from world-renowned facilities. This event has proven to be one of the most effective networking platforms in Asia, where biotechnology companies can converge to showcase their products and services, to promote concepts into market, to explore investment and collaborative opportunities, and to create new international biotechnology alliances [95, 96].

The creation of new biotechnology research facilities in India has also coincided with increased training initiatives as evidenced by the development of numerous

hands-on training programs and workshops in the areas of biotechnology and bioinformatics under themes such as “Wet Lab-Dry Lab integration.” The International Centre for Genetic Engineering and Biotechnology (ICGEB) conducts training workshops for up to 60 students, postdoctoral fellows, and scientists three or four times annually. In addition, another 60 doctoral students benefit from an association between the ICGEB and Jawaharlal Nehru University (JNU), New Delhi. Key institutes such as Aravinda Biosolutions Pvt. Ltd [97] and the Institute of Bioinformatics and Applied Biotechnology (IBAB) [98] also offer projects and courses in bioinformatics and cheminformatics. Many leading biotechnology companies in India such as Biocon India, Nicholas Piramal, Ranbaxy, Astrazeneca, Ocimum Biosolutions, and DNA Research Centre are actively involved in bioinformatics, vaccine, and drug screening, and genome/proteome research. Global leaders in biotechnology and life sciences such as Agilent Technologies have also shown considerable interest in the introduction of more genomics- and proteomics-related technology into India’s growing life sciences market [99]. In premier academic institutes like IIT, bioinformatics is now being taught alongside other, more traditional subjects, and various modules at the postgraduate levels dealing with biological computing are being approved by the appropriate governing bodies. In such institutes, the main interest is in structural and functional genomics, and their contributions to the scientific community in India and abroad is measured by the development of sequence and structural analysis services and metaservers. Recently, an important alliance was announced in India between the companies Ariadne Genomics and Jubilant Biosys, which will allow people from both companies to access databases of mammalian signaling and disease pathways [100]. Jubilant Biosys has developed many useful resources, which include: *PathArt* for molecular interaction networks from curated databases; *Kinase ChemBioBase*, a comprehensive database of small molecules for research on kinase targets; and *GPCR Ligand Database*, a small-molecule ligand database [101].

Premier Biosoft, a company in central India, is involved in the development of algorithms and software suites for high-throughput genome analyses. The major products developed by this company include SimVector (vector cloning and DNA analysis tool), Array Designer (high-throughput PCR primer and probe designing tool), and Beacon Designer (high-throughput design of real-time probes like TaqMan Probes, FRET Probes, etc.). There are many other companies that are also developing tools for genome-wide sequencing and for designing cross-species molecular probes and primers. The core facilities from leading universities across the United States are major customers, and they use these tools for in-house custom development of primers and probes in order to reduce any dependence on relatively costly professional suppliers of oligomers.

In recent years, pharmaceutical companies have expressed an interest for the development of target drugs. Therefore, related areas like Computer-Aided Drug Design (CADD), effective methods for Quantitative Structure Activity Relationship (QSAR), molecular modeling, and eventually structural biology are becoming key areas for bioinformatics and proteomics research in India. However, groups at premier institutes like CDRI, IGIB, and IISC are also involved in developing new

algorithms for gene expression analysis and artificial intelligence techniques to interpret biological data, which parallels the bioinformatics research that is occurring in developed countries. Recently, there has been interest in developing databases that contain diversified information about biological systems at the species level as well as a network of complex cellular processes such as the Cancer Cell Map (<http://cancer.cellmap.org/cellmap/>). The database of Quantitative Cellular Signaling is a repository of models of signaling pathways <http://doqcs.ncbs.res.in/> (<http://doqcs.ncbs.res.in/>); Conformation Angles Database (CADB), Bangalore (<http://cluster.physics.iisc.ernet.in/cadb/>), is another potentially useful database. Various research institutes in India are also involved in the development of databases such as the International Legume database and information services (ILDIS), which is being developed at NBRI (<http://www.nbri-lko.org/>).

Institute of Bioinformatics (IOB), International Tech Park Ltd., Bangalore, India, was founded by Akhilesh Pandey, a professor at Johns Hopkins University, USA. Pandey and colleagues from Bangalore and Baltimore manually curated functional annotation of human X chromosome and identified 696 protein-coding X-linked genes [102]. Researchers from IOB, India, along with researchers from Johns Hopkins and University of Wurzburg, Germany, have taken a great leap toward developing an interactome map for humans [103]. More than 25,000 human PPIs were mapped based on information generated from several databases, such as the Human Protein Reference Database (HPRD), Biomolecular Interaction Network Database (BIND), and Database of Interacting Proteins (DIP). Interestingly, comparison of 25,464 human PPIs with 16,069 yeast, 5625 worm, and 24,587 fly PPIs revealed only 16 PPIs common in all four organisms but revealed 42 PPIs common to human, worm, and fly [104]. The identification of protein and gene interactions sets a platform for the system biology studies [104]. Another initiative of IOB, creation of human plasma protein database to functionally annotate the subproteomes of human plasma, was featured on the cover of the journal *Proteomics* [105] and the *Journal of Proteome Research* appreciated the initiatives of IOB [106]. Even though most of India's success stories described above have been in "non-plant" areas, it is evident that India can take on a significant leadership role in the application of bioinformatics in plant proteome research.

Opportunities in Functional Proteomics/Interactomics

As described above, proteome-level research may be categorized as discovery proteomics or functional proteomics. India and other developing nations can play an important role in functional proteomics research by elucidating the activities regulated by proteins identified in discovery proteomics research. Even without the most sophisticated or costly laboratory equipment, comprehensive structural and functional characterization of proteins identified in high-throughput laboratories can be performed in the universities and research institutes in India and other developing nations perhaps using existing infrastructure. As more information is generated about the identified proteins, a thorough understanding of how different proteins interact with each other to facilitate necessary cellular activities (i.e., interactomics) will almost certainly develop. Furthermore, with increased information and knowledge being generated from various proteomics projects, comprehensive databases with archived protein sequences,

expression profiles, structural features, and functional activities will need to be created and maintained, which is another area in which developing nations, such as India, can more actively participate. Bioinformatics can also play a crucial role in the creation of effective algorithms to predict protein structure that may further facilitate research into protein structure and function [103]. Assuming the availability of technological advances, India can play a major role in cheminformatics research, the development of data resources, and data analysis techniques in genomics and proteomics research areas. India has started realizing the importance of proteomics research, and recently the President of India, Dr. A.P.J. Abdul Kalam, stated that “India has the potential to tap research opportunities in proteomics and biochips that can help understand biological processes and treat diseases. This is possible even though the country has missed the opportunity to partner in the human genome project” [107]. Dr. Kalam emphasized that the Indian biomedical community should build partnerships in the ongoing Proteomics Project.

47.3 CONCLUSIONS

The complete annotation of the human genome was accomplished through the collaborative efforts of numerous researchers from laboratories around the world in the publicly and privately funded Human Genome Project. Similarly, the genome sequences of plants and other organisms have also been completed or are currently in progress. The amount of information generated from these collaborative efforts bodes well for the burgeoning field of proteomics research if a similar spirit of cooperation is adopted. In terms of the discovery aspects of proteomics, laboratories in leading research institutions in India and other developing nations have the capabilities to perform protein separations using techniques like 2-DGE and LC. Often, researchers in these institutions rely on collaborations with their partners in other parts of the world to achieve these objectives. However, not all researchers in developing nations currently have access to researchers in developed countries to perform these types of experiments. Akhilesh Pandey and colleagues exemplified how a team effort among researchers from Bangalore (Institute of Bioinformatics) to Baltimore (John Hopkins University) could benefit proteomics projects, and this initiative was praised by an editorial published in *Nature Genetics* entitled “A feat worth replicating” [108]. Indeed, it must be replicated in the area of plant proteome research! What nations like India and other developing countries possess that developed countries must make better use of is a vast and perhaps untapped research force. A total population of over one billion in both India and China constitutes one of their greatest assets. India alone produces over 2.5 million university graduates annually, 650,000 postgraduates and 1500 PhDs in the engineering and life sciences fields alone [109], but training and infrastructure are required in order to effectively perform proteomics-based research in developing countries such as India. To adequately address the challenges that will be faced in terms of feeding the rapidly expanding global population, the developing world must rise to the challenge, and the developed nations must take full advantage of the increased interest and enthusiasm displayed by researchers from all countries.

47.4 FIVE-YEAR VIEWPOINT

The infrastructural requirements for functional proteomics endeavors are not as prohibitive as some of the instrumentation required for high-throughput discovery proteomics, and many laboratories in the developing world already have such capacities. In fact, many leading institutions in these countries are already contributing toward the area of functional proteomics; however, the efforts are largely fragmented due to individual researchers and institutional focus, which preclude large, concerted efforts. For example, if all the leading institutions in all nations—developed as well as developing—were to create a research consortium whose mandate would be to characterize every protein that has been implicated in mediating abiotic and biotic responses in plants, the progress that could be achieved would, undoubtedly, be significant. Of course, a collaborative venture of this nature would require unprecedented cooperation between participating institutions and a seamless transfer of information between research facilities. It would require a higher level of thinking that goes beyond individual research goals. Developing nations and their research facilities are capable of contributing toward this particular goal, and the participation of laboratories from the developed world will enhance these types of efforts. To avoid duplication of efforts and to generate useful information, there is a need to form an international committee on proteomics research, which could serve as a platform to raise awareness, distribute resources in the developing and developed world, and bring stakeholders together to focus on their common interest, which ideally would be the improvement in the quality of life for everyone.

ACKNOWLEDGMENTS

Financial assistance from the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Alberta Agricultural Research Institute (AARI), and the University of Alberta for the work from the Kav laboratory described in this chapter is gratefully acknowledged. Valuable suggestions on the status of bioinformatics in India from Ms. Pooja Jain, University of Nottingham, UK, is acknowledged.

REFERENCES

1. Wilkins MR, et al. 1995. *Biotechnol Genet Eng Rev* **13**:19–50.
2. Bradshaw RA and Burlingame AL. 2005. *IUBMB Life* **57**:267–272.
3. Palcy S and Chevet E. 2006. *Proteomics* **6**:5467–5480.
4. Hattan SJ, et al. 2005. *J Proteome Res* **4**:1931–1941.
5. Li X. 2005. *Proteomics* **5**:3423–3441.
6. O'Farrell PH. 1975. *J Biol Chem* **250**:4007–4021.
7. Smith JA, et al. 2006. *Mol Plant Microbe Interact* **19**:150–160.
8. Kersten B, et al. 2002. *Plant Mol Biol* **48**:133–141.
9. Ong SE and Mann M. 2005. *Nat Chem Biol* **1**:252–262.

10. Hanash SM. 2001. *Curr Opin Mol Ther* **6**:538–545.
11. Pirondini A, et al. 2006. *Chromatogr B Anal Technol Biomed Life Sci* **833**:91–100.
12. Zolla L, et al. 2003. *Plant Physiol* **131**:198–214.
13. Heazlewood JL, et al. 2004. *Plant Cell* **16**:241–256.
14. Unlu M, Morgan ME, and Minden JS. 1997. *Electrophoresis* **18**:2071–2077.
15. Tonge R, et al. 2001. *Proteomics* **1**:377–396.
16. Kislinger T, et al. 2005. *J Am Soc Mass Spectrom* **16**:1207–1220.
17. Dunkley TP, et al. 2004. *Biochem Soc Trans* **32**:520–523.
18. Turecek F. 2002. *J Mass Spectrom* **37**:1–14.
19. Guerrero IC and Kleiner O. 2005. *Biosci Rep* **25**:71–93.
20. Ong SE, et al. 2002. *Mol Cell Proteomics* **1**:376–386.
21. Blagoev B, et al. 2003. *Nat Biotechnol* **21**:315–318.
22. Gygi SP, et al. 1999. *Mol Cell Biol* **19**:1720–1730.
23. Anderson L and Seilhamer J. 1997. *Electrophoresis* **18**:533–537.
24. Agrawal GK and Rakwal R. 2006. *Mass Spectrom Rev* **25**:1–53.
25. Ping P. 2003. *Circ Res*. **93**:595–603.
26. Monti M, et al. 2005. *Biosci Rep* **25**:45–56.
27. Ito T, et al. 2001. *Proc Natl Acad Sci USA* **98**:4569–4574.
28. Li S, et al. 2004. *Science* **303**:540–543.
29. Formstecher E, et al. 2005. *Genome Res* **15**:376–384.
30. Giot L, et al. 2003. *Science* **302**:1727–1736.
31. Rual JF, et al. 2005. *Nature* **437**:1173–1178.
32. Kung LA and Snyder M. 2006. *Nat Rev Mol Cell Biol* **7**:617–622.
33. Uttamchandani M, Wang J, and Yao SQ. 2006. *Mol Biosyst* **2**:58–68.
34. Kersten B and Feilner T. 2006. *Methods Mol Biol* **355**:365–378.
35. Agrios GN. 1997. Significance of Plant Diseases. In: *Plant Pathology*, 4th edition, Academic Press, San Diego, CA.
36. Wang W, Vinocur B, and Altman A. 2003. *Planta* **218**:1–14.
37. Yamaguchi T and Blumwald E. 2005. *Trends Plant Sci* **10**:615–620.
38. FAO. 2005. Global network on integrated soil management for sustainable use of salt-affected soils. FAO Land and Plant Nutrition Management Service, Rome, Italy; <http://www.fao.org/ag/agl/agll/>
39. Borojevic S. 1990. *Principles and Methods of Plant Breeding* (Borojevic S, ed.), Elsevier, Amsterdam.
40. Gelvin S. 2000. *Annu Rev Plant Physiol Plant Mol Biol* **51**:223–256.
41. Birch R. 1997. *Annu Rev Plant Physiol Plant Mol Biol* **48**:297–326.
42. Prakash C. 2001. *Plant Physiol* **126**:8–15.
43. Apse MP, et al. 1999. *Science* **285**:1256–1258.
44. Kav NNV, et al. 2004. *Ann Appl Biol* **145**:217–230.
45. Jain S, et al. 2006. *Plant Physiol Biochem* **44**:253–259.
46. Srivastava S, Fristensky B, and Kav NNV. 2004. *Plant Cell Physiol* **45**:1320–1324.
47. Srivastava S, et al. 2006. *Plant Biotech J* **4**:529–549.

48. Yajima W, Hall JC and Kav NN. 2004. *J Agric Food Chem* **52**:5063–5070.
49. Subramanian B, Bansal VK and Kav NN. 2005. *J Agric Food Chem* **53**:313–324.
50. Sharma N, et al. 2006. *Plant Sci* **172**:95–110.
51. Yajima W and Kav NN. 2006. *Proteomics* **6**:5995–6007.
52. Health Canada. 1994. *Guidelines for the Safety Assessment of Novel Foods*, Vol 2: *Genetically Modified Microorganisms and Plants* Government of Canada, Ottawa, ON, Canada.
53. Metzdorff SB, et al. 2006. *Plant Biol (Stuttg)* **8**:662–672.
54. Corpillo D, et al. 2004. *Proteomics* **4**:193–200.
55. Catchpole GS, et al. 2005. *PNAS* **102**:14458–14462.
56. Ruebelt MC, et al. 2006. *J Agric Food Chem* **54**:2169–2177.
57. Gregersen PL, Brinch-Pedersen H, and Holm PB. 2005. *Transgenic Res* **14**:887–905.
58. Baudo MM, et al. 2006. *Plant Biotech J* **4**:369–380.
59. Hajheidari M, et al. 2005. *Proteomics* **5**:950–960.
60. Askari H, et al. *Proteomics* **6**:2542–2554.
61. Dooki AD, et al. 2006. *Proteomics* **6**:6498–6507.
62. Gazanchian A, et al. 2007. *J Exp Bot* **58**:291–300.
63. Trisiriroj A, Jeyachok N, and Chen ST. 2004. *Proteomics* **4**:2047–2057.
64. Degrave WM, et al. 2002. *Appl Bioinformatics* **1**:53–56.
65. He F. 2006. *Proteomics* **6**:397–403.
66. Cui S, et al. 2005. *Proteomics* **5**:3162–3172.
67. Yan S, et al. 2005. *Proteomics* **5**:235–244.
68. Yan SP, et al. 2006. *Mol Cell Proteomics* **5**:484–496.
69. Shen S, Jing Y and Kuang T. 2003. *Proteomics* **3**:527–535.
70. Yang P, et al. 2006. *Proteomics* **6**:3330–3338.
71. Dai S, et al. 2006. *Proteomics* **6**:2504–2529.
72. Xie Z, et al. 2006. *Proteomics* **6**:474–486.
73. Zhao C, et al. 2005. *Proteomics* **5**:961–972.
74. Wang BC, et al. 2006. *Proteomics* **6**:2555–2563.
75. Wang H, et al. 2006. *J Exp Bot* **57**:827–835.
76. Wang Y, et al. 2005. *Proteomics* **5**:4496–4503.
77. Ramakrishnan C. 2001. In Memoriam: Professor G.N. Ramachandram (1922–2001). *Protein Sci* **10**:1689–1691.
78. Ramachandran GN and Kartha G. 1954. *Nature* **174**:269–270.
79. Ramachandran GN and Kartha G. 1955. *Nature* **176**:593–595.
80. Subramanian E and Ramachandran GN. 2001. *Nat Struct Biol* **8**:489–491.
81. Srinivasan R, Balasubramanian R, and Rajan SS. 1975. *J Mol Biol* **98**:739–747.
82. Ramakrishnan C and Soman KV. 1982. *Int J Pept Protein Res* **20**:218–237.
83. Chakravarthy S and Varadarajan R. 1999. *Structure* **7**:723–732.
84. Reddy CK, Das A, and Jayaram B. 2001. *J Mol Biol* **314**:619–632.
85. Bhinge A, et al. 2004. *Structure* **12**:1989–1999.
86. Bahadur RP, et al. 2004. *J Mol Biol* **336**:943–955.
87. De S, Krishnadev O, Srinivasan N, and Rekha N. 2005. *BMC Struct Biol* **5**:15.

88. Guharoy M and Chakrabarti P. 2005. *Proc Natl Acad Sci USA* **102**:15447.
89. Prakash T, et al. 2005. *J Mol Biol* **345**:937–955.
90. International Rice Genome Sequencing Project. 2005. *Nature* **436**:793–800.
91. Loftus B, et al. 2005. *Nature* **433**:865–868.
92. Jayaraman KS. 2005. *Nature* **436**:480–483.
93. India emerging biotech leader in APAC region. *Deccan Herald* newspaper, June 10, 2004. <http://www.deccanherald.com/deccanherald/jun102004/b3.asp>
94. No glitches for biotech in Bangalore. Biospectrum, the business of biotech. August 12, 2004. <http://www.biospectrumindia.com/content/BSTOP20/BSTop2004/10408121.asp>
95. Bangalore Bio 2006 gains momentum. Biospectrum, the business of biotech. May 12, 2006. <http://www.biospectrumindia.com/content/bioEvents/10605121.asp>
96. Bangalore gears up for Bangalore Bio 2006 Biospectrum, the business of biotech. March 15, 2006. <http://www.biospectrumindia.com/content/bioEvents/10603152.asp>
97. Live projects and training programs in bioinformatics, molecular modeling and drug designing at Aravinda Bio, Hyderabad. Bioinformatics India, November 20, 2006. <http://www.bioinformaticsindia.com/new/?q=node/726>
98. Institute of Bioinformatics and Applied Biotechnology (IBAB), September 14, 2005, Biospectrum, the business of biotech. <http://www.biospectrumindia.com/content/Bio-Special/10509143.asp>
99. Agilent technologies to introduce genomics technology into India's growing life sciences market. PALO ALTO, Calif., April 20, 2005. <http://www.agilent.com/about/newsroom/presrel/2005/20aprca05035.html>
100. Ariadne Genomics and Jubilant Biosys announce strategic alliance February 16, 2004. <http://www.ariadnegenomics.com/company/news/02162004.html>
101. Srinivasan N. 2006. *Bioinformation* **1**:105–109.
102. Harsha HC, et al. 2005. *Nat Genet* **37**:331–332.
103. Edwards D and Batley J. 2004. *Trends Biotechnol* **22**:232–237.
104. Gandhi TK, et al. 2006. *Nat Genet* **38**:285–293.
105. Muthusamy B, et al. 2005. *Proteomics* **5**:3531–3536.
106. Currents. 2006. *J Proteome Res* **5**:1041.
107. Focus on gene chips and proteomics, says Kalam. The Hindu, Business line Aug 06, 2005 <http://www.thehindubusinessline.com/2005/08/06/stories/2005080601180300.htm>
108. Editorial. 2005. *Nat Genet* **37**:329.
109. Peet NP. 2005. *Drug Discov Today* **10**:1137–1139.

INDEX

Note: Page numbers followed by *f* and *t* indicate figures and tables, respectively; *b* indicates boxed material.

- ABA-responsive protein ABR17, of *M. trunculata*
root, regulation by fungal pathogen, 228*t*, 236
- Abiotic stress(es)
cellular response to, proteomics of, 605–628
advances in (future directions for), 624–625
progress in, 607–620, 608*t*–609*t*, 624
defense response against, analysis,
transcriptomics and proteomics of,
621–623, 622*t*
differential effects of, proteomics of, 620–621
research on, progress in, 606–621
types of, 605, 606*f*
- ABP1. *See* Auxin-binding protein 1 (ABP1)
- ABPP. *See* Activity-based protein profiling (ABPP)
- ABPs. *See* Activity-based probes (ABPs)
- Acetylation, 42–43
- Acetyl-CoA synthetase, anther, regulation of, 257
- Acid invertase, anther, 254, 255*t*
- Acid phosphatase, in root tissue, regulation of,
226*t*, 233
- Acrylamide, alkylation with, 24
- Acrylates, 49, 50*t*
- Actin
microfilaments, 62
in nucleus, 335
pollen, 267, 270
- Actin-binding protein, pollen, 267
- Actin-depolymerizing factors, pollen, 267
- Activity-based probes (ABPs), 57, 160
based on reversible inhibitors/ligands, 52–53,
53*f*, 54*t*
directed, using irreversible inhibitors, 48–52,
50*t*, 51*t*, 52*f*
- Activity-based protein profiling (ABPP)
probe for, 48*f*
strategies for, 47–53
- Activity-based proteomics, 48
- Acyloxymethylketone, 49, 50*t*
- Adenylate kinase, anther, regulation of, 256
- ADP/ATP translocator, in etioplasts, 354*f*, 355
- ADP-glucose pyrophosphorylase (AGPase), anther,
regulation of, 257
- Affinity-based probes, 48, 48*f*
directed, using irreversible inhibitors, 48–52,
50*t*
- Affinity purification, in study of protein–protein
interactions, 67–68
- AGP. *See* Arabinogalactan proteins (AGP)
- Agrobacterium tumefaciens*, cell transformation
methodology using, 71–72
- AIR9 (microtubule-binding protein), 280*t*, 284
- Albumins, 196
in mature seed, 193*f*, 197
- Aldolase
C-1, anther, regulation of, 257
in nucleus, 334
- Alfalfa. *See* *Medicago sativa*
- ALIS (Alexa-fluor-labeled internal protein
standard), 111
- Allergens
in food, proteomics of, 215, 415
nut, proteomics of, 415
oleosins as, 415
pollen, analysis of, 270–271
seed, proteomics of, 415
wheat, proteomics of, 215

- Alternative hypothesis (H_1), 130, 131*f*
- Aluminum hydroxide, in metal oxide affinity chromatography, for phosphopeptide and phosphoprotein enrichment, 425–426
- Aluminum oxide, in metal oxide affinity chromatography, for phosphopeptide and phosphoprotein enrichment, 425–426
- Alyssum lesbiacum*, cellular stress response, to heavy metals, 608*t*, 615
- AMAs. *See* Antibody microarrays (AMAs)
- AM fungus, plant symbiosis, root proteins regulated by, 226*t*, 233
- Aminoacyl-tRNA synthetase, in MT proteome analysis, 283
- Aminocyclopropane-carboxylate oxidase (ACC oxidase), in root tissue, regulation of, 231*t*
- Aminopeptidases, in plasma membrane, 323
- Amino-terminal propeptide (NTPP), 392, 404
- Ammonium assimilation, in etioplasts, 354*f*, 355
- Ampholine, 17, 18, 18*f*
- AMPS (2-acrylamido-2-methyl propane sulfonic acid), 20
- α -Amylase isoform III, in *planta*- and in *vitro*-derived, in rice, 87, 88*f*
- Amyloplast, wheat grain, proteome of, 215–219
- Angiosperms, seed formation in, 191–192
- Animal MT end-binding protein (EB1), plant orthologs, 281
- ANN. *See* Artificial neural network (ANN)
- Annexin, anther, regulation of, 257
- Anther
- development of, 249–250, 250*f*
 - proteome analysis of, 252–254, 255*t*–256*t*
 - proteomics of, 6
 - advances in (future directions for), 258–259
 - biotic and abiotic stresses and, 254–258
 - during development, 252–254, 255*t*–256*t*
 - flow chart for, 252, 253*f*
 - methodology for, 251–258
 - progress in, 250–251, 258
 - strategy for, 251–258
 - and transcriptomics, comparison of, 258
 - rice, proteomics of, and transcriptomics, comparison of, 258
 - wall layers of, 249–250, 250*f*
- Anther loculi, 249–250, 250*f*
- Anthozoa, GFP-like proteins, 70–71
- Antibody(ies), in detection of phosphorylation, 427
- Antibody microarrays (AMAs), for analysis of protein phosphorylation states in plant extracts, 494
- Antifungal protein R, in root tissue, regulation of, 230*t*, 232
- Antioxidant enzymes
- expression of, in rice leaves, 246
 - regulation of, environmental stress and, 234
- Aphanomyces euteiches*, root protein regulation by, 228*t*, 235–236
- Appressorium stage, of AM symbiosis, 633
- proteomics at, 636–637
- APX, anther, 252, 253
- regulation of, 257
- Aquaporins, in plasma membrane, 321–322
- Arabidopsis* (incl. *Arabidopsis thaliana*), 5
- BELL proteins, 578–579
- calcium sensors
- identification of, bioinformatics approaches for, 648–649
 - targets, identification of, 653–655, 654*f*
- calcium signaling networks, proteomics of, 645–658
- cell cultures (transient system), for studying protein localization, 73–74, 74*f*, 75*f*, 76*t*
- cell suspension culture, cell wall proteomics of, 302, 302*t*, 303*t*
- cellular processes, proteomics studies of, 152
- cellular stress response
- to excessive or inadequate light, 609*t*, 617–618
 - to gaseous pollutants, 608*t*, 616
 - to heat stress, 609*t*, 618
 - to heavy metals, 608*t*, 614–615
 - to low temperature, 609*t*, 619
 - to nutrient availability, 608*t*, 613
 - to salinity, 608*t*, 612
- cell wall
- proteome, 149, 149*f*
 - organ specificity of, 302, 302*t*
 - secretome, 149
- cell wall proteins
- acting on polysaccharides, 300–301, 301*f*
 - functional classification of, 300, 301*f*
 - oxido-reductase, 300–301, 301*f*
 - protease, 300–301, 301*f*
 - proteomics of, 296
 - advances in (future directions for), 304–306
 - in signaling, 301, 301*f*
- chloroplast
- proteins of
 - import of, 346
 - plastid-encoded, 346
 - post-translational modifications of, 347
 - targeting peptides, 346
 - proteomics of, 340, 341, 344, 346
 - advances in (future directions for), 348–349
 - progress in, 348
 - quantitative techniques for, 341, 347–348
- cob* mutant, 294

- COBRA (cell wall protein), functions of, 294
- cotyledons
- etiolated, peroxisomal proteins, 380
 - peroxisomes, isolation of, 380–382, 381*f*
- cultured cells, protein expression in, global analysis of, 147
- cytokinin-signaling pathway, analysis of, 579–580
- databases, Internet resources for, 144*t*
- defense response against abiotic stress, transcriptomics/proteomics of, 621–623, 622*t*
- defense response against pathogens, proteomics of, 588, 591*t*
- developmental processes, proteomics studies of, 152
- differential expression proteomics of, 152–153, 154*t*–157*t*
- ecotypes, leaf protein patterns in, 240
- EF-hand proteins in, 648*f*
- gene expression, 143–145
- genome, 166
- analysis, and microtubule-binding protein identification, 281
 - Internet resources for, 144*t*
 - sequencing of, 62
 - size of, 143
- genotypes, proteomics studies of, 152–153
- germinating embryos, defense against fungal infection, 200
- germination in, stored proteins and mRNA in, 201, 202*f*
- glyoxysomal proteome, characterization of, 383–385, 384*f*, 385*t*
- kinases, 55
- KNOX proteins, 578–579
- leaf
- peroxisomal proteome, characterization of, 385–386, 385*t*
 - proteomics of, progress in, 240
 - senescence, studies of, 243–244
- leucine-rich repeat receptor-like kinases, 469–470
- analysis, proteomics approaches to, 469–470
- light harvesting complex kinase, 55
- MADS box TFs, 578
- MAP65 family, 281
- membrane protein interactome, 574
- microtubule organization 1 (*mor1*) mutant, 281
- mitochondria
- isolation of, 364–365
 - proteomic analysis of, 144*t*, –145*t*, 149*f*, 150–151
 - purity, evaluation of, 365
 - subfractionation of, 365–366
- mitogen-activated protein kinase (MAPK), downstream signaling, 494–495, 495*f*
- mutants, proteomics studies of, 152–153
- MYB proteins, 578–579
- N*–myristoylome, 159
- nuclear matrix, proteomics of, 328
- nuclear proteins, 149–150, 149*f*, 328
- nuclear proteome, comparative analysis of, 335–336, 336*f*
- nucleolus
- proteomic analysis of, 144*t*, –145*t*, 150, 334
 - proteomics of, 328
- nucleus, proteomics of, 328
- oleosins, 408
- organelles
- with double membranes, proteomic analysis of, 144*t*, –145*t*, 149*f*, 150–151
 - with single membranes, proteomic analysis of, 144*t*, –145*t*, 149*f*, 151–152
- OVATE protein family, 578–579
- oxido-reductases, in cell wall, organ specificity of, 302, 303*t*
- peroxisomal genes, 378
- peroxisome, proteomics of, 144*t*, –145*t*, 149*f*, 151–152, 379–380
- advances in (future directions for), 387
 - progress in, 386
- plasma membrane proteomics, literature review, 310, 311*t*–312*t*
- plastid
- protein targeting, 355–356, 357*f*
 - proteomic analysis of, 144*t*, –145*t*, 149*f*, 150–151
- PMR6, functions of, 294
- pmr6* mutant, 294
- pollen
- collection of, 264
 - mature
 - functional classification of, 265–268, 266*t*, 268–270, 269*f*
 - reference maps, establishment of, 264–265
 - protein expression in, global analysis of, 147
 - proteins of, isoforms, 265
 - proteomics of
 - advances in (future directions for), 272
 - progress in, 264, 271–272
 - and transcriptomics, comparison of, 268–270, 269*f*
 - transcriptomics of, 261–262, 261–263, 263
 - and proteomics, comparison of, 268–270, 269*f*
- post-translational modifications in, MS studies of, 153–159

Arabidopsis (incl. *Arabidopsis thaliana*),
(Continued)

protein complexes
 BN-PAGE of, 566
 isolation, TAP-tagging technology for, 543–556
 protocol for, 551–552
protein kinases, 485
protein–protein interaction mapping for, 578–580
protein–protein interactions in, study of, 69, 144*t*–145*t*, 160, 574
proteins, 143–145
proteome, 143–164
 changes in response to abiotic stresses, 620
 proteomics of, 143–146, 379
 14–3–3, 519–522
 advances in (future directions for), 160–161
 methodology for, 146–147
 strategy for, 146–147
 topics of, 146, 146*f*
root
 adventitious formation, protein regulation in, 225
 gravity and, 233
root hair development
 plant hormones and, 232
 proteomic studies of, 231–232
rsh mutant, 294
salt tolerance, proteins in, 235
SDD1 cell wall protease, 305
secreted peptides, database of, 103
secretome analysis in, 84–85
seed
 albumins, 193*f*, 197
 germination
 metabolic control of, 203–204
 proteomics of, 198–199
 globulins (cruciferins), 197
 oil bodies
 enzymes, 411–412
 minor proteins, 412
 structural proteins, 410–411
 proteins, 193*f*
 expression, global analysis of, 144*t*, 147
 post-translational modifications, during germination, 199–200
 storage organs, 192
specialized cells, protein expression in, global analysis of, 144*t*–145*t*, 147–160
spiral 1 (*spr1*) mutant, 281
subcellular compartments, proteomics of, 144*t*–145*t*, 146*f*, 148–150, 149*f*
TALE homeodomain proteins, 578–580

tissues, protein expression in, global analysis of, 144*t*–145*t*, 147–160
trichome, protein expression in, global analysis of, 147
Tyr phosphorylation in, emerging evidence, 499–514
vacuoles
 proteomic analysis of, 144*t*, –145*t*, 149*f*, 151–152
 proteomics of, 393–394
 advances in (future directions for), 403–404
vacuome, 151
vpey mutant, vacuolar proteomics, label-free quantitative approach, 401–402, 402*f*
Arabinogalactan proteins (AGP), 296
 analysis of, 297–299
 in signaling, 301
 Yariv-affinity chromatography of, 296, 297
Arbuscules, 631, 632*f*
 definition of, 630
ARD, expression in root development, regulation of, 225, 229*t*
Arginine
 dimethylation, 43, 43*t*
 monomethylation, 43, 43*t*
ARK3, in plasma membrane, 322
Arsenic, root proteins regulated by, 227*t*, 234, 235
Artificial neural network (ANN), in prediction of plant protein localization, 64*t*, 65
L-Ascorbate peroxidase, expression in root development, regulation of, 230*t*, 232
Aspartate carbamoyltransferase, in nucleus, 335
Aspartate protease, in cell wall, 301
Aspartyl proteases
 activity-based probes for, based on reversible inhibitors/ligands, 53, 53*f*, 54*t*
 in planta- and *in vitro*-derived, in rice, 87, 88*f*
ATCase, in nucleus, 335
ATPases, in plasma membrane, 321–322
ATP sulfurylase, in root tissue, heavy metals and, 227*t*, 235
ATP-synthase, in root tissue, heavy metals and, 227*t*, 235
AtXTH4 xyloglucan endotransferase, in *Arabidopsis thaliana* cell wall, 304
Auxin
 and root formation in rice, 232
 root proteins regulated by, 230*t*, 232
Auxin-binding protein 1 (ABP1), expression in root development, regulation of, 225
Bacillus subtilis, secretome analysis in, 84, 89
Bacteria
 protein interaction mapping for, 577

- rice defense response against, differential proteome profiles during, 596–597
- Bacteroids, 631–632, 632*f*
 - B. japonicum*, proteomics of, 639–640
 - definition of, 630
- Band-pass (BP) emission filter, 112–113, 113*f*
- Band-sharpening effect, 14
- Barley
 - seeds
 - germinating, proteomics of, 199
 - proteomics of, 195–196
 - vacuoles, proteomics of, 394–395
- Barrel medic. *See Medicago truncatula*
- p*-Benzoquinone reductase, in root tissue, heavy metals and, 227*t*, 235
- Berberine-bridge enzyme
 - (S)-reticulin:oxido-reductases, in cell wall, 300
 - organ specificity of, 302–304, 303*t*
- Beta expansin
 - anther, 254, 255*t*
 - regulation of, 257
 - pollen, as allergen, 271
- Beta vulgaris* L. *See* Sugar beet
- Bimolecular fluorescence complementation (BiFC), in study of protein interactions *in vivo*, 76–77
- Bioinformatics, 4–5
 - and asynchronous flow of functional evidence vs. proteome-wide analysis, 505
 - in chloroplast proteomics, 340, 346–347
 - development of, 107–108
 - in gel-based proteomics, 107–125
 - advances in (future directions for), 123–124
 - in identification of putative plant protein Tyr kinases, 505–506, 507*f*
 - in India, 722–724
 - and microtubule-binding protein identification, 280–281
 - in MS-based proteomics, 127–139
 - advances in (future directions for), 138
 - in peptidomics, 97–98
 - in prediction of plant protein localization, 63, 66
- Biolithics, 71
- Bio-Lyte, 17, 18
- Biomarkers, peptide, 92
- 2,3-Bisphosphoglycerate-independent phosphoglycerate mutase, anther, regulation of, 257
- Black cottonwood. *See Populus trichocarpa*
- BLAST, 300
- Blue native-polyacrylamide gel electrophoresis (BN-PAGE), 557–570
 - of acetyl-CoA carboxylase, 566
 - advances in (future directions for), 567–568
 - advantages of, 559
 - applications of, 559, 560*f*
 - and BN PAGE (native), 565
 - of cytochrome c maturation complex, 566
 - development of, 559, 560–561
 - digitonin in, 560–561, 562*b*
 - first dimension, downstream applications of, 564*f*, 565
 - of formate dehydrogenase complex, 566
 - and IEF-SDS-PAGE (denaturing), 563–565
 - of isovaleryl-CoA dehydrogenase, 566
 - limitations of, 561*b*
 - methodology for, 561–563
 - of mitochondrial protein import apparatus, 566
 - of plant electron transport chains, 566
 - of plastidic ClpP protease complex, 566
 - of plastidic ω -3 saturases, 566
 - of plastid protein import apparatus, 566
 - progress in, 560–561, 567
 - of protein complexes, results, 566–567
 - and SDS-PAGE (denaturing), 563, 564*f*
 - second and third dimensions, 563–565
 - staining procedures for, 564*f*, 565
 - in study of protein–protein interactions, 67, 68–69
 - of thylakoid protein insertion system, 566
 - of tobacco plastid-encoded RNA-polymerase complex, 566
- BN-PAGE. *See* Blue native-polyacrylamide gel electrophoresis (BN-PAGE)
- Bowman-Birk trypsin inhibitor, in root tissue, regulation of, 231*t*, 232
- Brain, proteomics, 6
- Brassica napus*, seeds
 - oil bodies, proteomics of, 410–412
 - proteomics of, 195
- Brassinosteroid (BR) receptor, 471
- Brassinosteroid (BR) receptor complex, in plasma membrane, 322
- Brassinosteroid (BR) signaling, 471–472
- Brassinolide (BL)
 - and rice root development, 232
 - root proteins regulated by, 230*t*, 232
- BRET, for studying protein–protein interactions *in vivo*, 67, 76
- Buffer(s)
 - bifunctional, 20
 - conventional, 13
 - discontinuous, 13
 - Laemmli's, 13
 - O'Farrell lysis, 12
 - pI select, 20
 - soluble amphoteric, conventional isoelectric focusing (IEF) in, 17–19

- Buffer(s) (*Continued*)
 wash, for rice leaf protein extraction, 168–169, 169*f*
- CA. *See* Carrier ampholytes
- CA074, 51, 52*f*
- Cadmium
 and differential expression proteomics in *Arabidopsis*, 156*t*
 root proteins regulated by, 227*t*, 228*t*, 234, 235
- Caenorhabditis elegans*, interactome, 574, 575
- Caesalpinioideae, 180
- CA-IEF, 17, 19, 21
 development of, 17
 as preparative technique, 17, 19
 problems with, 19–20
- CA-IEF/IPG gels, mixed-type, 21–22, 28
- Calcium-binding proteins. *See also* Calcium sensors
 in plasma membrane, 322
 as pollen allergens, 271
- Calcium sensors, 646–647, 646*f*, 648*f*
 identification of, 648–652
 bioinformatics approaches for, 648–649, 651*f*
 high-throughput experimental approaches for, 650–652, 650*b*, 651*f*
- pollen, 267
- targets, 646–647, 646*f*
 identification of, 652–655
 by screening of expression libraries, 652, 653*f*
 using protein chips, 653–655, 654*f*
 by Y2H screens, 652–653
- Calcium signaling, phases of, 646–647, 646*f*
- Calcium signaling networks (in plants), 645–658
 deciphering, 656
 protein complexes in, analysis of, using TAP-tagging technology, 655
 protein components of, 648
 proteomics of
 advances in (future directions for), 656–657
 progress in, 656
- Calcium signature, generation of, 646–647, 646*f*
- Caleosin, 408, 411
- Callase, 262, 263*f*
- Calreticulin
 anther, regulation of, 256
 pollen, 267
- Camera(s), CCD, 112–113
- CaM-like protein, pollen, 267
- Capillary ZE-MS (CZE-MS), 17
- Capsicum annuum*, cellular stress response, to excessive or inadequate light, 609*t*, 618
- Carbohydrate esterases, in cell wall, 300
- Carbohydrate lyases, in cell wall, 300
- Carbonylation, of seed proteins, in germination, 199–200
- Carboxy-terminal propeptide (CTPP), 392
- Carrier ampholytes, 11–12, 28
 focusing, 19
 focusing properties of, 17
 in isoelectric focusing (IEF). *See* CA-IEF
 molecular mass distribution of, 17
 polydispersity of, 17
 production of, variability in, 20
- Caspase, in plasma membrane, 323
- Castor beans, seed, oil bodies, proteomics of, 410–412
- Catalase, pollen, 267
- Cation-exchange chromatography (CAX), for peptide extraction from samples, 95
- CAX. *See* Cation-exchange chromatography (CAX)
- CBB. *See* Coomassie Brilliant Blue (CBB)
- CCOMT, anther, regulation of, Δ 256
- CDA. *See* Covariant discriminant algorithm (CDA)
- CDTA, in cell-wall protein extraction, 297
- Celiac disease, 208
- Cell fractionation, in determination of subcellular protein localization, 66–67
- Cell-free expression systems. *See also* Wheat germ, cell-free protein synthesis system
 for eukaryotic proteins, 661–669
 advances in (future directions for), 667
 application to functional proteomics, 663–665, 664*t*, 665*t*, 666*f*
 progress with, 667
- Cell-mapping proteomics, 4
- Cell signaling, 6. *See also* Signal transduction
- Cellulase, pollen, 267
- Cellulose
 in cell wall, 293, 294*f*
 synthesis of, 276
- Cellulose synthase
 and microtubules, 276
 pollen, 267
- Cell wall
 components of, 293, 294*f*
 and disease resistance, 294–295
 lignins in, 293
 polysaccharides in, 293
 proteomics of, 293–308
 structure of, 293, 294*f*
 supramolecular assembly of, studies of, advances in (future directions for), 306
- Cell wall proteins (CWP)
 acting on polysaccharides, 300–301, 301*f*

- analysis, difficulties of, 295
- in culture medium, analysis, 296
- direct analysis of, 297, 298*f*
- extraction
 - calcium chloride solution for, 297
 - CDTA in, 297
 - destructive methods for, 297, 298*f*
 - difficulties of, 295
 - lithium chloride solution for, 297
 - mannitol solution for, 297
 - methodology for, 296–300
 - NaCl solution for, 297
 - nondestructive methods for, 297, 298*f*
 - protocols for, evaluation of, 300
 - strategies for, 296–300, 298*f*
- fractionation of, 297
 - strategies for, 298*f*
- functional classification of, 300–301, 301*f*
- functions of, 293–295
 - studies of, advances in (future directions for), 305–306
- glycosylated
 - analysis, 296
 - analysis of, 297–299, 299*f*
- glycosylation of, 302
- identification of, strategies for, 298*f*
- with interacting domains, 301, 301*f*
- labile, 293, 294*f*, 295
- of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 235
- miscellaneous class of, 301, 301*f*
- in MT proteome analysis, 283–284
- oxido-reductase, 300–301, 301*f*
- phosphorylation of, 302
- and plasma membrane proteins,
 - interactions/relationships of, 323
- post-translational modifications of, 302
- protease, 300–301, 301*f*
- proteomics of, 296
 - advances in (future directions for), 304–306
 - applications of, 300–304
 - methodology for, 296–300
 - progress in, 304
 - results, 300–304
 - strategies for, 296–300
- in signaling, 301, 301*f*
- strongly bound to matrix, 293, 294*f*, 295
- structural, 293, 294*f*, 301*f*, 302
- in subproteomes, 296
- types of, 295
- weakly bound to matrix, 293, 294*f*, 295
- Cell-wall remodeling, pollen proteins in, 266*t*, 267, 269, 269*f*
- as allergens, 271
- Cesium, and differential expression proteomics in *Arabidopsis*, 156*t*
- Cetyltrimethylammonium bromide (CTAB)
 - precipitation, 86
- Chaperones. *See also* Heat shock proteins
 - anther, 252
 - pollen, 267
- Chaperonin, pollen, 267
- Chemical ionization (CI), 34
- Chemical proteomics, 47–59
 - advances in (future directions for), 58
- Chemical reactivity, 48
- Chemometrics, 118–119
- Chickpea
 - nuclear proteins
 - in DNA replication and transcription, 334–335, 334*f*
 - enrichment, assessment of, 330
 - extraction, 329
 - functional classification of, 332–335, 334*f*
 - in gene regulation, 334, 334*f*
 - identification of, 330
 - isolated, purity assessment, 329–330, 330–332, 331*f*
 - metabolism-related, 334*f*, 335
 - miscellaneous class, 334*f*, 335
 - post-translational modifications of, 332
 - in protein degradation, 334*f*, 335
 - in protein folding, 334*f*
 - proteomics of, progress in, 335–336
 - in signaling, 334, 334*f*
 - structural, 334*f*, 335
 - in translation, 334*f*, 335
 - in transport, 334, 334*f*
 - two-dimensional reference map, construction of, 332, 333*f*
 - unknown, 332, 334*f*
 - nucleus
 - isolation of, 329
 - proteomics of, 328–338
 - advances in (future directions for), 336–337
 - reference proteome map for, 328
- Chitinase, post-translational modifications of, 302
- Chlamydomonas*, light harvesting complex kinase, 55
- Chlamydomonas reinhardtii*
 - chemical proteomics of, 55, 56–57, 57*f*
 - phosphorylation sites in, 55
- Chlorophyll, biosynthesis of, 356

- Chloroplast(s), 39, 215–216
 compartments of, 339, 340*f*
 proteomics of, 341
 in defense and stress responses, proteomics of, 599–600
 formation of, 351–352
 functions of, 339, 344
 genome of, 339
 ligand affinity chromatography proteomics of, 341
 proteins of, 339
 extraction of, 342, 342*f*, 343
 functional classification of, 334
 mass spectrometry analysis of, 342, 342*f*, 344
 post-translational modifications of, 347
 separation of, 342, 342*f*, 343–344
 subcellular localization, 344–345
 proteomics of, 6
 advances in (future directions for), 348–349
 applications of, 344–348
 challenges of, 339–340
 methodology for, 341–344, 342*f*
 progress in, 341, 348
 quantitative, 341, 347–348
 results, 344–348
 sample preparation for, 342–343, 342*f*
 strategies for, 341–344, 342*f*
 Triton-insoluble fraction, 342*f*
 ultrastructure of, 339, 340*f*
- Chromatin, 335
- CI. *See* Chemical ionization (CI)
- CID. *See* Collision-induced dissociation (CID)
- CLASP1, in MT proteome analysis, 283
- Clinostat rotation, and differential expression
 proteomics in *Arabidopsis*, 155*t*
- CMS. *See* Cytoplasmic male sterility (CMS)
- COBRA protein, post-translational modifications of, 302
- Cold acclimation, cellular response to, proteomics of, 609*t*, 618–620
- Cold acclimation-specific protein CAS18, of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 235–236
- Cold-induced anther protein, OsCIA, 256
- Collision-induced dissociation (CID), 37, 40
 with MS3, for phosphopeptide fragmentation, 430
- Continuous-flow cell-free (CFCF) translation system, 661–669
- Coomassie Brilliant Blue (CBB), 12
 micellar, 26
 sensitivity in protein detection, 251, 252*f*
- Copper-binding proteins, in root tissue, 235
- Corn, 5
- Cotyledons, 191, 192
 dicot, 192
 electron microscopy of, 378*f*
 monocot, 192
- Covariant discriminant algorithm (CDA), in
 prediction of plant protein localization, 64*t*, 65–66
- Cowpea
 cellular stress response, to heavy metals, 608*t*, 615
 symbiotic relationships, proteomic analysis of, 639
- Crop improvement
 need for, 716–717
 proteomics in, 716–719
- Crop plants, conventional breeding *versus* genetic engineering of, 717
- Cruciferins, 197
- CTPP. *See* Carboxy-terminal propeptide (CTPP)
- CWP. *See* Cell wall proteins (CWP)
- Cyan fluorescent protein (CFP), 75
- Cyanine dyes, 27–28, 111–112, 112–114
- Cyclophilin, pollen, 267
- Cysteine
 dyes targeting, 28
 saturation labeling, 28
- Cysteine proteases
 in cell wall, 301
 inhibitor, epoxysuccinyl-type activity-based
 probes derived from, 51, 52*f*
 irreversible inhibitors, as directed affinity-based
 probe, 49, 50*t*
- Cysteine proteinase inhibitor, anther, regulation of, 256
- Cysteine synthase, anther, regulation of, 256
- Cytochrome c oxidase, subunit VB, anther,
 regulation of, 256
- Cytochrome P450, in root tissue, heavy metals
 and, 227*t*, 235
- Cytoplasmic male sterility (CMS), in rice,
 257–258
- Cytoskeletal proteins, 62
 pollen, 266*t*, 267, 269, 269*f*
 protein interaction mapping for, 580
- Cytoskeleton, 62
 plasma membrane and, 323
- CZE-MS. *See* Capillary ZE-MS (CZE-MS)
- Database(s)
 Arabidopsis, 152, 160
 Internet resources for, 144*t*
 endosymbiont genomes, 634
 EST, from plants, 380
 of phosphorylation sites, 428

- of plant plasma membrane proteome, 311*t*–312*t*, 319
- plant protein membrane, 144*t*
- for plant proteomics, 129–130
- predictive tools, 77
- PROSITE, 300
- of protein families, 648
- protein interaction, 145*t*
- protein localization data, 77
- proteomics, 77
- rice, 171
- of rice anther proteome, 254
- rice cDNA, 172
- searching, 128–135, 129*f*
 - performance achieved, 134–135
- Soybean EST, protein identification using, 384*f*, 385*t*, 386
- specialized, 144*t*
- SUBA, 346
- subproteome, 144*t*–145*t*
- TIGR Soybean Gene Index, 383
- Data integration, 129–130
- Data mining, in peptidomics, 97–98
- DE. *See* Delayed ion extraction (DE)
- Deep Purple, 26–27, 113
- Deep Purple total protein stain (DPTPS), 26–27
- Defense proteome analysis
 - in abiotic stress response. *See also* Abiotic stress(es)
 - transcriptomics and proteomics of, 621–623, 622*t*
- advances in (future directions for), 602–603
- data analysis in, 595
- differential proteomics using 2D-DIGE for, 594–595
- LCM for, 593–594, 598, 599*f*
- methodologies for, 589–595
- PEG fractionation studies for, 592, 593*f*, 594*f*
- progress in, 587–588, 590*t*–591*t*, 601
- protein identification in, 595
- rationale for, 587–588
- sequential extraction methods for, 592–593
- strategies for, 588, 589–595, 589*f*
- subcellular, 599–601
- Dehydration stress, 200
- Delayed ion extraction (DE), 36
- De novo* peptide sequencing, 40, 41*f*, 42, 135–138
 - current approaches, 135–137, 136*f*
 - exhaustive search, 137
 - existing tools, 137
 - genome search, 137
 - performance achieved, 137
- Density gradient centrifugation, cell fractionation and, 67
- Dephosphorylation, 485
 - pollen proteins mediating, 267
- Desiccation tolerance, 200
- Deubiquinating enzymes (DUBs), 444
 - analysis of, 450–451
- Developing countries, proteomics in, 713–730
- 2-DGE. *See* Two-dimensional (2D) gel electrophoresis (2-DGE)
- Diacylglycerol acyltransferase, in maize oil bodies, 412
- Dicots. *See also Arabidopsis*
 - embryo of, 192
 - leaves, veins of, 239
 - pollen, collection of, 263–264
 - roots of, 223, 224*f*
 - seeds
 - proteomics of, 192–195
 - structure of, 193*f*
- Difference in-gel electrophoresis (DIGE), 27–28, 96, 111
 - 2D-, for differential proteomics of plant defense response, 594–595
- Differential peptide display, 98–99
- DIGE. *See* Difference in-gel electrophoresis (DIGE)
- Digitonin, 560–561, 562*b*
- Disaccharide translocators, in plasma membrane, 321
- Discovery proteomics, 714–716
- Disease-resistance response protein pi 49, of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 236
- Disk-electrophoresis, 11
- Dormancy, seed, 191
 - in *Arabidopsis*, 198–199
 - hormonal control of, 192
- DPTPS. *See* Deep Purple total protein stain (DPTPS)
- Drosophila melanogaster*, interactome, 574, 575
- Drought, cellular responses to, proteomics of, 607–610, 608*t*
- DSKs. *See* Dual specificity kinases (DSKs)
- DSPs. *See* Dual specificity phosphatases (DSPs)
- Dual specificity kinases (DSKs), 500
- Dual specificity phosphatases (DSPs), 501
- DUBs. *See* Deubiquinating enzymes (DUBs)
- Dynamins, in plasma membrane, 323
- Dynein, 281
- E-64, 51, 52*f*
- EB1, plant orthologs, 281
- ECD. *See* Electron capture dissociation (ECD)
- Edman sequencing, 251–252, 428
- EES. *See* Excitation-emission spectroscopy (EES)

- EF. *See* Elongation factor(s)
- EF-hand proteins, 647, 648 *f*
 as pollen allergens, 271
- EI. *See* Electron ionization (EI)
- EigenMS, 137
- Electrokinetic methodologies, 11
- Electron capture dissociation (ECD), for
 phosphopeptide fragmentation, 430
- Electron ionization (EI), 34
- Electron microscopy (EM), of plant cells, 61–62
- Electron transfer dissociation (ETD), for
 phosphopeptide fragmentation, 430
- Electron transport chains (plant), BN-PAGE of, 566
- Electrospray ionization (ESI), 1, 33
 cell fractionation and, 67
 development of, 34–35, 36
 nano- (nESI), 36
 Q-TOF, combined, 37, 41
 time of flight (TOF) analyzer coupled to, 37
- Electrospray ionization (ESI)-Q-TOF, in
 proteomics, 96
- Elongation factor(s)
 EF-1 α , 282
 EF-1 β ', expression in root development,
 regulation of, 230*t*, 232
 expression in root development, regulation of,
 230*t*, 232
 in MT proteome analysis, 283
 in plasma membrane, 323
 pollen, 267
- Embryo
 dicot, 192
 monocot, 192
 plant, 191
 structures of, 192
- Endo-1,3- β -D-glucosidase, in root tissue,
 regulation of, 231*t*, 232
- Endopeptidase(s), in sample preparation, 39, 39*t*
- Endoplasmic reticulum/Golgi pathway
 nonclassical, in plants, 85, 88
 proteins of, quantitative proteomics analysis of,
 84
- Endoproteases, in vacuolar proteomics, 399
- Endosperm, 191
 triploid phase, in angiosperms, 192
 wheat grain
 gliadins, 207–208
 glutenin subunits, 207–208
 glutens, 207
 non-gluten proteins of, 208
 protein fractions, characterization of, 214–215
 proteins of
 analysis of, 208–210
 extraction of, 208–210, 209 *f*
 fractionation of, 208–210
 functions of, 208
 proteome of, 207–222
 characterization, methodology for, 210
 and cultivar identification, 214
 environment and, 213–214
 during grain development, 210–213, 213 *f*
 temperature and, 213–214
- Endothecium, anther, 249–250, 250 *f*
- Endotransglycosylate MERI-5, in plasma
 membrane, 323
- Enolase, anther, 253
- Entrez, 130
- Envelope, chloroplast, proteomics of, 342 *f*, 343,
 344–345
- Environmental proteomics, 153, 154*t*–157*t*
- Environmental stress, and proteome, 227*t*, 228*t*,
 234–236
- Enzyme activity, multisite protein phosphorylation
 and, 436–437
- Epicocconone, 26–27
- Epidermis
 anther, 249–250, 250 *f*
 leaf, 239, 240 *f*
 proteomics of, 241–242
- Epoxy succinyl moiety, 49, 50*t*
- ER. *See* Endoplasmic reticulum/Golgi pathway
- Erysiphe cichoracearum*, infection of *Arabidopsis*
 thaliana, 294
- ESI. *See* Electrospray ionization (ESI)
- ESTs. *See* Expressed sequence tags (ESTs)
- ETD. *See* Electron transfer dissociation (ETD)
- Ethylene
 and differential expression proteomics in
 Arabidopsis, 157*t*
 root proteins regulated by, 228*t*, 232
- Etioplast(s)
 heterotrophic, metabolic pathways of, 354–355,
 354 *f*
 intact, isolation of, 352, 353 *f*
 light-induced conversion to chloroplasts,
 351–352
 proteome dynamics during, 356–357, 358 *f*
 photosynthesis-dependent autotrophy, 351–352
 proteome dynamics, during light-induced
 chloroplast development, 356–358, 358 *f*
 proteomics of, 351–360
 advances in (future directions for), 359
 progress in, 358–359
 strategies for, 352–353
 purification of, 352, 353 *f*
 structure of, 351
- Eukaryotes, proteome, complete description of,
 2–3

- Excitation-emission spectroscopy (EES), in chemical proteomics, 55
- Exopolysaccharonase, pollen, 267, 269
- Expansins
in cell wall, 300
as pollen allergens, 271
- Expressed sequence tags (ESTs), 62, 713
Triticum, 208
- Expression proteomics, 2, 4
- Extensin-like protein, pollen, 270
- Extensins, in cell wall, 293, 297
functions of, 294
- Extracellular matrix (ECM), in defense and stress responses, proteomics of, 600–601
- False negatives (FN), 130–131
- False positives (FP), 131
- Far-Western blotting, in study of protein–protein interactions, 67, 69
- Fast atom bombardment (FAB), 33, 34–35
- FCA (nuclear protein), 334
- Ferguson plots, 12
- Ferredoxin/thioredoxin redox system, in etioplasts, 355
- Fibrillins, in chloroplast, 345
- FLIM. *See* Fluorescence lifetime imaging microscopy (FLIM)
- Flooding, cellular stress response to, proteomics of, 608*t*, 612–613
- Flowering plants. *See* Angiosperms
- Fluorescence lifetime imaging microscopy (FLIM), 75
- Fluorescence resonance energy transfer (FRET)
in chemical proteomics, 55
for studying protein–protein interactions *in vivo*, 67, 74–77
- Fluorescence spectral microscopy (FSPIM), 75
- Fluorescent stain, 26
- Fluorescent tagging of full-length proteins (FTFLP), 72–73
- Fluoromethylketones, 49, 50*t*
- Fluorophosphonates, 49, 50*t*
- FluoroProfile, 27
- Formate dehydrogenase, activity, multisite protein phosphorylation and, 436
- Fourier transform–ion cyclotron resonance (FTICR), 38, 41, 44
- fragile fibre2* mutant, 282
- Free-response operator characteristics (FROC)
curves, 114–115, 115*f*, 120
- FRET. *See* Fluorescence resonance energy transfer (FRET)
- Fructokinase, anther, 254, 255*t*
- Fructose 1,6-bisphosphatase (FBPase), 355
- Fructose bisphosphate aldolase-like protein, in root tissue, regulation of, 229*t*, 233
- FSPIM. *See* Fluorescence spectral microscopy (FSPIM)
- FTFLP. *See* Fluorescent tagging of full-length proteins (FTFLP)
- FTICR. *See* Fourier transform–ion cyclotron resonance (FTICR)
- Functional modules, 572, 573*b*
- Functional proteomics, 48
opportunities in, 724–725
vs. discovery of novel targets, 714–716
- Fungal infection, defense against, in germinating embryos, 200
- Fungi, rice defense response against, differential proteome profiles during, 595–596
- Fungus. *See also* Symbioses, AM
ectomycorrhizal, genomics, 634
pathogenic
genomics, 634
secretome analysis in, 84
plant-colonizing, genomics, 634
Fusarium graminearum, secretome analysis in, 85
- Fusion proteins, 70
- Galectins, activity-based probes for, based on reversible inhibitors/ligands, 53, 53*f*, 54*t*
- Gametophyte, male. *See* Pollen
- GAPDH
as MT-binding protein, significance of, 285
in MT proteome analysis, 283, 284
in nucleus, 332, 334–335
- Gaseous pollutants, cellular stress response to, proteomics of, 608*t*, 616
- Gel gradient electrophoresis, 14
- GenBank dbEST, 633, 634
- Genomics, 166, 166*f*
comparative, 574
- Germination. *See* Seed, germination
- Germin-like protein
pollen, 269
in root tissue, regulation of, 231*t*, 232
- Germins, in cell wall, 300
- GFP. *See* Green fluorescent protein (GFP)
- Gliadins, 207–208
wheat, as allergen, 215
in wheat amyloplasts, 217
- GLIP1, functional analysis of, 85, 89
- Globulins, 196
in mature seed, 196–197
- Glomales intraradices*, plant symbiosis, root proteins regulated by, 226*t*, 233
- Glomales mosseae*, plant symbiosis, root proteins regulated by, 226*t*, 233

- Glucan 1,3- β -glucosidase precursor, *in planta*- and *in vitro*-derived, in rice, 87, 88*f*
- Glucose 6-phosphate, in etioplast, 354*f*, 355
- Glutamine synthetase/glutamate synthase (GS/GOGAT) cycle, in etioplasts, 354*f*, 355
- Glutathione peroxidase, in root tissue, heavy metals and, 227*t*, 235
- Glutathione *S*-transferase (GST)
 anther, 252, 255*t*
 expression in root development, regulation of, 227*t*, 228*t*, 230*t*, 231*t*, 232, 235
 pollen, 267
S-Glutathionylation, in *Arabidopsis*, 159
- Glutelins, 196
- Glutenins, 207–208
 in wheat amyloplasts, 217
- Gluten proteins, 207
- N*-Glycans, 455
 chemically or enzymatically released, analysis of, 460–463, 463*b*
 functions of, 457–458
 in plants
 possible structures of, 456, 457*f*
 structural analysis of, 460–463, 463*b*
 structural analysis of, 455, 460–463, 463*b*
- Glycine hydroxymethyltransferase, in root tissue, heavy metals and, 228*t*, 235
- Glycine max*. *See also* Soybean
 cotyledons
 peroxisomal proteins, characterization of, 384*f*, 385*t*
 peroxisomes, isolation of, 381*f*, 382
 ESTs, 380, 387
 peroxisomal proteins, characterization of, 384*f*, 385*t*, 386
 proteomics of, 183–184, 185*f*, 187
 seeds, proteomics of, 194, 194*f*
 symbiotic relationships, proteomic studies of, 184
 tissues/organs, protein extraction protocols for, 180, 180*t*
- Glycine-rich protein(s), anther, 252
- Glycine-rich RNA-binding proteins (GRPs), in nucleus, 334
- Glycolytic enzymes
 interactions with microtubules, 285
 in nucleus, 333, 334
- Glycopeptides, analysis of, 463–465
- Glycoproteins
 glycoforms of, 455
 hydroxyproline-rich (HRGP), in cell wall, analysis of, 297–299, 299*f*
 identification of, 459–460
 macroheterogeneity of, 455
 microheterogeneity of, 455
- Glycosidases, suicide substrate, 51, 51*t*
- Glycoside hydrolases, in cell wall, 300
- Glycoside transferases, in cell wall, 300
- Glycosylation
 of cell wall proteins, 296, 302
 of leaf proteins, 244–245
- N*-Glycosylation
 analysis of
 advances in (future directions for), 466
 glycomic, 455
 glycoproteomic, 455
 by mass spectrometry, 458–459, 464*f*
 progress in, 465–466
 biological significance of, 455
 process of, 456
 of proteins in plants, 455–468
- Glyoxylate cycle, 377
- Glyoxysomes, 378, 378*f*
- Golgi complex, proteins of, 84
- GPI-anchored proteins (GAPs), *Arabidopsis*, 159, 319–320, 321*f*
- G-proteins
 in plasma membrane, 322, 323
 RAB, in plasma membrane, 322, 323
- Grapevines, cellular stress response, to herbicides, 608*t*, 616
- Graph-theoretical network analysis
 progress in, 571–572
 terminology for, 573*b*
- Gravitropism, 392
- Gravity, root proteins regulated by, 228*t*–229*t*, 233
- Green alga. *See Chlamydomonas reinhardtii*
- Green fluorescent protein (GFP), for protein localization assays, 70
- Green fluorescent protein (GFP)–protein fusions
 cell transformation methodology using, 72–73, 74*f*, 75*f*
 generation of, 72
 overexpression of, 72–73
- Growth, regulation of, multisite protein phosphorylation and, 436
- GTP-binding protein
 β -subunit, expression in root development, regulation of, 225, 229*t*
 in plasma membrane, 322
 RAN, in root tissue, regulation of, 231*t*, 232
- GTP-mediated signaling, pollen proteins in, 267
- Guanine nucleotide-binding protein beta subunit, in root tissue, heavy metals and, 227*t*, 235
- Gymnosperms, pollen tube, proteomics of, 272

- H-9, 53*f*
 chemical probe for kinases based on, 55
- Haustoria, 630
- Heat shock protein(s) (HSP), in plasma membrane, 323
- Heat shock proteins
 anther, 252, 253
 regulation of, 257
 HSP70/HSP40, in MT proteome analysis, 283, 284
 of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 235
 in MT proteome analysis, 284
 pollen, 267
- Heat stress
 cellular stress response to, proteomics of, 609*t*, 618
 and rice, proteomics studies of, 176–177
- Heavy metals
 cellular stress response to, proteomics of, 608*t*, 614–616
 root proteins regulated by, 227*t*, 228*t*, 234, 235
- Hemicellulose, in cell wall, 293, 294*f*
- Herbicide stress, cellular stress response to, proteomics of, 608*t*, 616–617
- Hidden Markov model (HMM), 137
 Novo, 42
 in prediction of plant protein localization, 64*t*, 65
- Histone H2B.2, expression in root development, regulation of, 225, 229*t*
- Histones, 335
- HMM. *See* Hidden Markov model (HMM)
- Holm oak, leaf
 cellular stress response, to drought, 608*t*, 610
 proteomics of, 240–241
- Homology modeling, 678–682, 679*f*
- Housekeeping protein(s), anther, 252
- H protein, of glycine cleavage system, anther, regulation of, 256
- Humans
 cellular protein composition, determination of, 67
 nucleolus, characterization of, 67
 proteomics, 5–6
- Hypocotyl, 192
 etiolated, cell wall proteomics of, 296, 302, 302*t*, 303*t*
- Hypothesis testing, 130–132, 131*f*
- IAC. *See* Inhibitor affinity chromatography (IAC)
- ICAT. *See* Isotope-coded protein label (ICAT)
- IEF. *See* Isoelectric focusing (IEF)
- iHop, 130
- IMAC (immobilized metal-ion affinity chromatography), 43
 for phosphopeptide and phosphoprotein enrichment, 424–425, 473–475, 474*f*
 in studies of post-translational modifications in *Arabidopsis*, 158
- Immobiline, 20, 29
 basic, 20
 carboxyl, 20
 pKs
 temperature and, 20
 urea and, 20
- Immobilized pH gradients (IPGs), 22–23
 in isoelectric focusing (IEF), 17
 isoelectric focusing (IEF) in, 19–21
 in peptide separation, 22
 pH ranges, 29
 porosity, 29
 principles of, 20
 in proteome analysis, 21, 21*f*
 re-formulation of, 29
 in rice proteomics studies, 169–170, 172–174
 in two-dimensional (2D) mapping, 22
- Immuno-based detection, 1
- Immunophilins, in chloroplast lumen, 346
- Immunoprecipitation, in study of protein–protein interactions, 67–68
- India
 bioinformatics in, 722–724
 proteomics research in, 719, 720–721
- Inhibitor(s)
 irreversible. *See* Ligand(s), irreversibly binding
 reversible. *See* Ligand(s), reversibly binding
- Inhibitor affinity chromatography (IAC), 48
 protein enrichment by, 52
- Initiation factor(s)
 in MT proteome analysis, 283
 in plasma membrane, 323
- InsPecT, 137
- The Institute for Genomic Research. *See* TIGR
- Insulin, processing, 96
- Interacting proteins, in plasma membrane, 322
- Interactome
 fly, 574, 575
 human, 574, 575
 worm, 574, 575
 yeast, 574
- Interactome walking, 572, 580
- Interactomics, 557, 571
 opportunities in, 724–725
- International protein index (IPI), 129
- International Rice Research Institute (IRRI), 165
- Internet resources, for proteomics studies, 144*t*–145*t*

- InterProScan, 300, 648
 Iodoacetamide, alkylation with, 24
 Ion-Cyclotron, 96
 Ion detectors, 38
 Ionization source, 34–35, 34*f*
 Ion trap, 37
 IPGs. *See* Immobilized pH gradients (IPGs)
 IRGSP, 165–166, 171
 IRRI. *See* International Rice Research Institute (IRRI)
 ISG15, 444, 451
 Isoelectric focusing (IEF), 11–12, 17–22
 conventional, in soluble amphoteric buffers, 17–19
 in immobilized pH gradients (IPGs), 19–21
 in leaf proteomics, 241
 Isoelectric point (pI), 11–12
 assessment of, advances in (future directions for), 29–30
 Isoelectrofocusing (IEF), in two-dimensional (2D) mapping, 12
 Isoforms, in mature pollen protein analysis, 265
 Isoliquiritigenin 2-*O*-methyltransferase, of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 235
 Isomerases, in plasma membrane, 323
 Isotope-coded affinity analysis, for
 phosphoproteomics, 473–475, 474*f*–475*f*
 Isotope-coded protein label (ICAT), 153

 JA hormone
 and endophytic colonization in rice, 234
 protein expression induced by, in rice, 232
 root proteins regulated by, 231*t*, 232
 Jun activation domain-binding protein, anther, 252, 255*t*

 Katanin-like MT-severing protein, 282
 Kinase interaction sequence (KIS), 508
 Kinases
 activity-based probes for, based on reversible inhibitors/ligands, 53, 53*f*, 54*t*
 in nucleus, 334
 photoaffinity labeling, with H-9 derivative, 55–56, 56*f*
 plant
 functions of, 55
 molecular tools targeting, development of, 53–57, 56*f*, 57*f*
 in plasma membrane, 322
 pollen, 267
 subfamilies, in plants, identification of, 55–58
 Kinesin-like calmodulin-binding protein (microtubule-binding protein), 280*t*

 Kinesin-like CaM-binding protein (KCBP), mutant, 282
 Kinesin-related proteins, in *Arabidopsis*, 281
 Kinesins
 in MT proteome analysis, 284
 in plasma membrane, 323
 KIS. *See* Kinase interaction sequence (KIS)

 LADH, anther, regulation of, 257
 Laemmli's buffer, 13
 Laser scanners, 112–114
 Late embryogenesis abundant (LEA) proteins
 Arabidopsis, 158
 expression in root development, regulation of, 230*t*, 232
 Latex allergen-like proteins, in root tissue, heavy metals and, 227*t*, 235
 LB-TT solution, 168–169, 168*f*, 169*f*
 Leaf
 functions of, 239
 proteins of, extraction techniques for, 241
 proteomes, post-translational modifications in, 244–245
 proteomics of, 6, 239–248
 advances in (future directions for), 247
 progress in, 240–241, 247
 results, poor reproducibility, 241
 techniques for, 241–242
 rosette, cell wall proteomics of, 296, 301, 302, 302*t*, 303*t*
 senescence, and proteomics, 242–244
 structure of, 239, 240*f*
 vascular system of, 239, 240*f*
 proteomics of, 242
 veins of, 239, 240*f*
 Lectins
 activity-based probes for, based on reversible inhibitors/ligands, 53, 53*f*, 54*t*
 with curculin-like domains, in *Arabidopsis thaliana* cell wall, 304
 post-translational modifications of, 302
 in root tissue, regulation of, 226*t*, 233
 Legumes
 nucleus, proteomics of, 328–338
 phylogeny of, 180
 protein extraction protocols for, 180, 180*t*
 proteome of, 179–189
 proteomics studies of, 181–186
 advances in (future directions for), 187
 progress in, 186–187
 reference proteome map for, 328
 root proteins, regulation by fungal pathogen, 228*t*, 235–236
 seeds, proteomics of, 192–194, 194*f*

- subfamilies of, 180
- symbiotic relationships, proteomic studies of, 182–183, 184, 187
- tissues/organs, protein extraction protocols for, 180, 180*t*
- Lesion mimic mutants (LMMs), defense response against, proteomics of, 591*t*, 598
- Leucine-rich repeat receptor-like kinases (LRR RLKs), 469–484
- analysis
 - advances in (future directions for), 481–482
 - Gateway-compatible construct kit for, generation of, 477–478
 - isotope-coded affinity approaches for, 473–475, 474*f*–475*f*
 - plant membrane proteomics and, 473
 - progress in, 481–482
 - proteomics approaches to, 470–471
- biological function of, 469–470
- BRI1
 - biochemical properties of, 478–480, 479*f*
 - in vivo* phosphorylation sites, identification of, by LC-MS/MS, 480
 - phosphorylation sites, functional characterization of, 480–481
- BRI1/BAK1, 470
 - activity of, 471–472
 - analysis of, 471–472, 472*f*
 - structure of, 471–472, 472*f*
- functional analysis of, by mutant
 - complementation, 476–477
- functions of, 471–472
 - biochemical characterization of, 477
- heterodimerization of, 471–472
- phosphoproteomics of, isotope-coded affinity analysis for, 473–475, 474*f*–475*f*
- phosphorylation of, 471–472
- phosphorylation sites, identification of, by LC-MS/MS, 476
- LHC. *See* Light harvesting complex (LHC)
- Ligand(s)
 - irreversibly binding, 48–49, 50*t*
 - linked to reporter tag, 48, 48*f*
 - reversibly binding, 48
 - labeling strategies using, 52–53, 53*f*, 54*t*
 - as recognition units, 52–53
- Light
 - and differential expression proteomics in *Arabidopsis*, 154*t*–155*t*
 - excessive or inadequate, cellular stress response to, proteomics of, 608*t*–609*t*, 617–618
- Light harvesting complex (LHC)
 - kinases, 55
 - proteins, *C. reinhardtii*, phosphorylation sites in, 55
- Lignins, in cell wall, 293
- Likelihood ratio, 134
- Lipase, in seeds, OB-associated, 411–412
- Lipid transfer protein (LTP)
 - anther, regulation of, 254, 256*t*
 - wheat, as allergen, 215
- Liquid-crystal chromatography, alignments and comparisons, 138
- Liquid-phase methodologies, 28–29
- Liver, proteomics, 6
- LMK. *See* Localization motif knowledgebase (LMK)
- LMMs. *See* Lesion mimic mutants (LMMs)
- Localization motif knowledgebase (LMK), in
 - prediction of plant protein localization, 63, 64*t*
- Long-pass (LP) emission filter, 112–113, 113*f*
- Lotus japonicus*
 - proteomics of, 183, 186, 187
 - root nodules, protein complexes, BN-PAGE of, 566
 - symbioses, proteomics of, 633
 - tissues/organs, protein extraction protocols for, 180, 180*t*
- Low molecular weight (LMW) proteomics, 92. *See also* Peptidomics
- LRR-receptor protein kinases, in signaling, 301
- LRR RLKs. *See* Leucine-rich repeat receptor-like kinases (LRR RLKs)
- lrr1* mutant, root protein regulation in, 225, 229*t*
- Lumen, chloroplast, 339, 340*f*
 - proteomics of, 342*f*, 345–346
- Lupinus albus*, cellular stress response
 - to drought, 608*t*, 609
 - to nutrient availability, 608*t*, 614
- Lutefisk, 137
- Lysine
 - acetylation, 43, 43*t*
 - dimethylation, 43, 43*t*
 - monomethylation, 43, 43*t*
 - trimethylation, 43, 43*t*
 - ubiquitination, 43, 43*t*
- Lysine minimal labeling, 27–28
- Macrosporangia, 249–250, 250*f*
- Magnaporthe grisea*, rice defense response against, differential proteome profiles during, 590*t*, 595–596
- Maize
 - adventitious root formation, proteomic studies of, 225–231
 - cellular stress response

Maize (*Continued*)

- to drought, 607–610, 608*t*
 - to excessive or inadequate light, 608*t*, 617
 - to gaseous pollutants, 608*t*, 616
 - to heavy metals, 608*t*, 615–616
 - to salinity, 608*t*, 611
 - cell wall proteins, proteomics of, 296
 - defense response against abiotic stress,
 - transcriptomics/proteomics of, 622*t*, 623
 - defense response against pathogens, proteomics of, 588, 591*t*
 - endosperm, proteomics of, 196
 - germinating embryos, defense against fungal infection, 200
 - lrt1* mutant, root protein regulation in, 225, 229*t*
 - oil bodies, enzymes in, 412
 - root mutant, proteomic studies of, 225
 - rtcs* mutant, root protein regulation in, 225, 229*t*–230*t*
 - shoot-borne root initiation, proteomic studies of, 225
 - tan1* mutant, 281
 - xylem sap proteome, 296
- MALDI. *See* Matrix-assisted laser desorption ionization (MALDI)
- MALDI–TOF mass spectrometry (MS), 16, 36, 41, 41 *f*, 251–252
- MALDI–TOF–TOF mass spectrometry, in peptidomics, 97, 99–100
- Male gametophyte. *See* Pollen
- MAPK. *See* Mitogen-activated protein kinase(s) (MAPK)
- MAPKK. *See* Mitogen-activated protein kinase kinase(s) (MAPKK)
- MAP65 protein family, 281, 285
- functions of, 282
 - isolation of, 282
 - in MT proteome analysis, 284
- MAPs. *See* Microtubule-associated proteins (MAPs)
- MAR-binding filament-like protein 1 (MFPL), 335
- MARs. *See* Matrix attachment regions (MARs)
- Mascot, 130, 210, 252, 595
- performance achieved, 134–135
 - scoring function, 133
- Mass analyzer, 34, 34 *f*, 36–38
- tandem-in-time, 37
 - time of flight (TOF), 37
 - types of, 36
- Mass detector, 34, 34 *f*
- MassLynx, 252
- Mass spectrometer
- components of, 34–35, 34 *f*
 - state-of-the-art, qualities of, 34

Mass spectrometry (MS), 2, 251–252. *See also*

- Capillary ZE-MS (CZE-MS)
 - advances in (future directions for), 44
 - alkylating step in, 25
 - bioinformatics and, 127–139
 - cell fractionation and, 67
 - database search, 128–135
 - existing tools for, 130
 - hypothesis testing, 130–132
 - direct analysis, 138
 - for identification of post-translational modifications, 42–44, 153–159
- MALDI-based, 1
- time of flight (TOF). *See* MALDI–TOF mass spectrometry (MS)
- in peptidomics, 95–97
- for qualitative analysis, 96
 - for quantitative analysis, 96
 - for sequence analysis, 96
- principles of, 33
- protein identification with, 40–44
- in proteomics, 33–44, 251–252
- quantitative methods, 2
- sample preparation for, 39, 39*t*
- tandem (MS/MS), 34, 37, 127, 128
- for peptide fragmentation identification, 40–41
 - in peptidomics, 96
- uses of, 33–34
- variations of, 5
- workflow of, 34 *f*
- Mass-to-charge (*m/z*) ratio, 33, 128
- Matrix-assisted laser desorption ionization (MALDI), 1, 33, 35–36, 138
- cell fractionation and, 67
 - development of, 34–35
 - matrix compounds for, 35–36, 35*t*
 - principle of, 35
 - time of flight (TOF) analyzer coupled to, 37
 - time of flight (TOF) MS. *See* MALDI–TOF mass spectrometry (MS)
- Matrix attachment regions (MARs), 335
- Matrix metalloproteinase (MMP), 102
- MCF. *See* Mitochondrial carrier family (MCF)
- MCP. *See* Microchannel plate (MCP) detectors
- Medicago sativa*
- cell wall proteins, proteomics of, 296
 - proteomics of, 184–186, 186 *f*, 187
 - tissues/organs, protein extraction protocols for, 180, 180*t*
- Medicago trunculata*, 5
- microsomal root proteins, subcellular proteomics of, 183
 - peptidome, 103

- proteomics of, 180–183, 181, 187
- root nodule development, proteomics of, 638
- root proteins, regulation by fungal pathogen, 228*t*, 235–236, 636
- seeds
 - desiccation tolerance in, 200
 - germination, metabolic control of, 204
 - proteomics of, 192–194, 194*f*
- skl* mutant, 638
- somatic embryogenesis in, proteomics of, 201–203
- sunm* mutant, 637
- symbiotic relationships, proteomic studies of, 182–183, 226*t*, 233, 633–636
- tissues/organs, protein extraction protocols for, 180, 180*t*
- Meiosis, in pollen development, 262, 263*f*
- Mesophyll, 239, 240*f*
- palisade cells, 239, 240*f*
- proteomics of, 242
- spongy cells, 239, 240*f*
- Mesorhizobium*, plant cell transformation
 - methodology using, 72
- Messenger RNA (mRNA), microarrays, 4
- Metabolic enzymes, in MT proteome analysis, 279*f*, 283, 284
 - significance of, 285
- Metabolism
 - in etioplasts
 - anabolic, 354–355, 354*f*
 - catabolic, 354–355, 354*f*
 - proteins involved in
 - in chloroplast, 345
 - in MT proteome analysis, 279*f*, 283, 284
- Metabolomics, 103, 138, 166, 166*f*
- Metalloproteases, activity-based probes for, based on reversible inhibitors/ligands, 53, 53*f*, 54*t*
- Metal oxide affinity chromatography (MOAC), for phosphopeptide and phosphoprotein enrichment, 425–426, 426*f*
- Metal oxide/hydroxide affinity chromatography (MOAC), in studies of post-translational modifications in *Arabidopsis*, 158
- Methionine, oxidation, 43, 43*t*
- Methylation, 42–43
 - in *Arabidopsis*, 159
- Methylmalonate-semialdehyde dehydrogenase (MMSDH), expression in root development, regulation of, 232
- MFP. *See* Multifunctional protein (MFP)
- MFP1. *See* MAR-binding filament-like protein 1 (MFP1)
- Microchannel plate (MCP) detectors, 38
- Microgametogenesis, 249–250, 250*f*
- Microspore(s), production of, 249–250, 250*f*, 262, 263*f*
- Microsporogenesis, in angiosperms, 249–250, 250*f*, 262, 263*f*
- Microtubule(s), 62, 275–276
 - disassembly of, 285
 - formation of, 276
 - functions of, 275
 - interactions with microtubule-binding proteins, 276–277
 - interphase cortical array, 276
 - metaphase spindle, 276
 - organization of, 276–277
 - mutants impaired in, 277, 281–282
 - phragmoplast, 276
 - plant hormones and, 276
 - pre-prophase band, 276
 - proteome of, dynamics of, 285–286, 286*f*
 - stress response and, 276, 285
- Microtubule-associated proteins (MAPs), 275–276. *See also* MAP65 protein family; Microtubule-binding proteins
 - classical, 276, 280
 - functions of, 276, 280, 280*t*
 - MAP70, 280*t*
 - MAP190, 282
 - nonclassical, 276, 280
 - interactions with microtubules, significance of, 284–285
 - in MT proteome analysis, 284
 - structural
 - identification of, 277
 - mutant screens for, 277
- Microtubule-binding proteins, 275–290. *See also* Microtubule-associated proteins (MAPs)
 - bundling activities, 282
 - classes
 - identified in MT co-assembly study, 279*f*
 - identified in tubulin-affinity chromatography, 279*f*, 282–283
 - co-assembly and co-purification of, using endogenous and neuronal microtubules, 277–278, 282
 - functions of, 280, 280*t*
 - identification of
 - genome analysis and, 281
 - informatics approaches, 280–281
 - in large-scale MT proteome studies, 279*f*, 283–284
 - in situ* localization experiments, 284
 - interactions with microtubules
 - dynamics of, 285–286, 286*f*
 - spatial, 285–286, 286*f*
 - temporal, 285–286, 286*f*

- Microtubule-binding proteins, (*Continued*)
- mutant screens for, 277
 - plant-specific, 280, 280*t*
 - genetic screens for, 281–282
 - identification of, 281–282
 - predicted, 280
 - proteomics of
 - advances in (future directions for), 287
 - informatics approaches, using plant genome sequence information, 277
 - methodologies for, 277–280
 - progress in, 286–287
 - strategies for, 277–280
 - purification of
 - biochemical approaches, 277–280
 - using MT and tubulin-affinity chromatography, 278–280, 279*f*, 282
 - regulation of, 276–277
 - in regulation of cortical MT orientation, 282
 - RNA-binding activity, 283, 285
- Microtubule-interacting proteins, 276
- Middle layer, anther, 249–250, 250*f*
- Mimosoideae, 180
- Mitochondria (plant)
- compartments of, 361–362
 - functions of, 361–362
 - cristae membrane structure of, 362
 - in defense and stress responses, proteomics of, 599–600
 - genome of, 362
 - hydrophobic proteome
 - analysis, improvement of, 371–373, 371*f*
 - characterization, progress in, 369–371, 370*t*
 - proteomics of, novel techniques for, 367*f*, 372–373
 - targeted proteomics of, by data mining, expression analyses, 372
 - inner membrane of, 362
 - integral membrane proteins
 - gel-based resolution of, 366–368
 - 2D 16-BAC PAGE/SDS-PAGE technique for, 367*f*, 368
 - Tricine double SDS-PAGE for, 366–368, 367*f*
 - non-gel-based LC-based resolution of, 368–369
 - RP-HPLC-MS/MS for, 368–369
 - intermembrane space of, 361
 - membrane protein fractions, preparation of, 366
 - metabolic characteristics of, 362
 - origin of, 362
 - outer membrane of, 361
 - protein complexes, BN-PAGE of, 566
 - proteins of
 - functional classification of, 334
 - targeting of, 362
- proteome
- dynamic characteristics of, 363–364
 - “global” profiling of, 363
 - oxidative stress and, 363–364
 - tissue-type differences in, 363–364
- proteomics of, 6, 362
- advances in (future directions for), 373–374
 - methodology for, 364–369
 - progress in, 363–364, 373
 - results, 369–373
 - strategy for, 364–369
 - structure of, 361–362
 - subfractionation of, 365–366
- Mitochondrial carrier family (MCF), 370–371, 370*t*
- Mitogen-activated protein kinase(s) (MAPK), plant
- downstream signaling, 494–495, 495*f*
 - signaling cascade, 494–495, 495*f*
- Mitogen-activated protein kinase kinase(s) (MAPKK), 500
- Mitosis, in pollen development, 262, 263*f*
- Mlo protein, in plasma membrane, 322
- Model
- biological, 707–708
 - definition of, 698*t*
- Modeling
- approaches, for systems biology, 707–708
 - definition of, 698*t*
- Molecular biology, 1
- Molecular mass (M_r), 12
- determination of, 12, 14, 15*f*
 - porosity gradients and, 14
 - resolution of, 16–17
 - versus* migration, plot of, 16
- Molecular recognition, 48
- Monoclonal antibody(ies) (MAb), 1
- Monocots. *See also* Rice
- embryo of, 192
 - leaves, veins of, 239
 - pollen, collection of, 263–264
 - roots of, 223, 224*f*
 - seeds
 - proteomics of, 195–196
 - structure of, 193*f*
- Mono-DHAR, anther, regulation of, 256
- Monosaccharide translocators, in plasma membrane, 321
- MOR1, in MT proteome analysis, 284
- Motor proteins, in MT proteome analysis, 279*f*, 284
- MS. *See* Mass spectrometry (MS)
- MtBcp1, in root tissue, regulation of, 226*t*, 233

- Mth1, in root tissue, regulation of, 226*t*, 233
- MTs. *See* Microtubule(s)
- MudPIT, 149
- in peptidomics, 93, 100
 - in vacuolar proteomics, 397–398
- Multi-copper oxidases, in cell wall, 300
- organ specificity of, 302, 303*t*
- Multifunctional protein (MFP), peroxisomal, 280*t*, 283
- MT-binding activity, 283
 - RNA-binding activity, 283, 285
- Mycorrhiza, mature, proteomics in, 634–636
- N*–Myristoylation, in *Arabidopsis*, 159
- m/z*. *See* Mass-to-charge (*m/z*) ratio
- α -NAC. *See* α -Nascent polypeptide-associated complex (α -NAC)
- NADH-dependent oxidoreductase, expression in root development, regulation of, 230*t*, 232
- NADPH oxidase, in plasma membrane, 323
- α -Nascent polypeptide-associated complex (α -NAC), in root tissue, salt stress and, 227*t*, 235
- Nascent polypeptide associated complex α chain, expression in root development, regulation of, 230*t*, 232
- NCBI nr database, 129
- NEDD8, 444, 451
- Network analysis. *See* Graph-theoretical network analysis
- Nickel, root proteins regulated by, 234
- Nicotiana tabacum*. *See* Tobacco
- Nitric oxide (NO), and leaf protein modifications, 245
- S*–Nitrosylation
- in *Arabidopsis*, 159
 - of leaf proteins, 245
- NMR spectroscopy
- protein structure determination by, 671, 675–677
 - and X-ray crystallography, for protein structure determination, comparison of, 678
- Nod factors, 226*t*, 233, 631, 639
- Nostoc commune*, cellular stress response, to excessive or inadequate light, 608*t*, 617
- NovoHMM, 137
- NS-134, 51
- NTPP. *See* Amino-terminal propeptide (NTPP)
- Nucleic acid binding protein, anther, regulation of, 257
- Nucleocytoplasmic shuttling, regulation of, 334
- Nucleolus
- human, characterization of, 67
 - plant, 62
 - characterization of, 67
- Nucleotide diphosphate kinase (NDPK), anther, regulation of, 255*t*, 257
- Nucleus (plant), 327–338. *See also* Chickpea, nucleus
- complexity of, 327
 - proteins of, functions of, 327–328
 - proteome, 327–328
 - comparative analysis of, 335–336, 336*f* - proteomics of
 - advances in (future directions for), 336–337
 - progress in, 335–336, 336*f* - structure of, 327
- Null hypothesis (H_0), 130, 131*f*
- Nutrition
- cellular stress response to, proteomics of, 608*t*, 613–614
 - and differential expression proteomics in *Arabidopsis*, 155*t*
- OB. *See* Oil bodies (OB)
- Octopus instrument, 19, 29
- O'Farrell lysis buffer (LB), 12
- Oil bodies (OB), 407–418
- enzymes in, 407, 408
 - functions of, investigation of, 412–413
 - proteins of, 407, 409, 409*f*
 - alkaline, separation and identification of, 409–410 - contaminant, 408
 - in developing and germinating seeds, studies of, advances in (future directions for), 414–415
 - functions of, 414, 414*f*
 - highly hydrophobic, separation and identification of, 409–410
 - post-translational modifications of, 412
 - structural, 407–408
 - targeting of, 412–413
 - proteomics of
 - advances in (future directions for), 414–415
 - progress in, 413–414, 414*f* - purification of, 408–413
 - seed, proteome of, 410–412
 - structure of, investigation of, 412–413
- Oilseed rape. *See* *Brassica napus*
- Oleosins, 410–411, 412–413
- and allergy, studies of, advances in (future directions for), 415
 - functions of, 407–408
 - insertion within oil bodies, structural proteomics of, 413
 - tissue distribution of, 408

- Oleosomes. *See* Oil bodies (OB)
- Omics, 166, 166*f*
- OMSSA, 130
- performance achieved, 135
- OPPP. *See* Oxidative pentose phosphate pathway (OPPP)
- Orbitrap, 38, 41, 96
- ORFeomes, 576–577
- Organelle(s). *See also specific organelle*
- proteins of, 84
- proteomics of, 378–379
- targeting sequences and localization, predictive tools, 145*t*
- Oryza sativa* L. *See also* Rice
- cellular stress response
- to drought, 608*t*, 610
- to flooding, 608*t*, 612–613
- to gaseous pollutants, 608*t*, 616
- to salinity, 608*t*, 610–611
- as model for functional genomics of cereal crops, genome sequence, 166, 167*f*
- ssp. Japonica and Indica, genome sequence, 166
- Osg6B, anther, regulation of, 254
- Osmotin protein homologue, anther, regulation of, 256
- Oxidative pentose phosphate pathway (OPPP), in etioplast, 354–355, 354*f*
- Oxidative stress, 377
- defense against, 234
- and differential expression proteomics in *Arabidopsis*, 154*t*
- plant mitochondria and, 363–364
- Oxidoreductases
- cell-wall, 300–301, 301*f*
- in cell wall, organ specificity of, 302, 303*t*
- p190 (microtubule-binding protein), 280*t*
- P. meyeri*, pollen tube, proteomics of, 272
- PAGE. *See* Polyacrylamide gel electrophoresis (PAGE)
- Papilionoideae, 180
- clades of, 180
- tribes of, 180
- Paralogs
- definition of, 265
- in mature pollen protein analysis, 265
- Particle bombardment, cell transformation using, 71
- Pathogen elicitor, and differential expression proteomics in *Arabidopsis*, 157*t*
- Pathogenesis-related protein
- 1, in root tissue, regulation of, 230*t*, 232
- Prb 1, in root tissue, regulation of, 231*t*, 234
- Pathogenesis-related protein(s) (PR proteins)
- as allergens, 271
- PR10, of *M. trunculata* root, regulation by fungal pathogen, 228*t*, 236
- PR-10a, in root tissue, regulation of, 231*t*, 232
- RSOsPR10, in root tissue, regulation of, 231*t*, 234
- Pathogens, and differential expression proteomics in *Arabidopsis*, 156*t*
- Pattern recognition, 118–119
- PCA. *See* Protein fragment complementation assays (PCA)
- PDH complex (PDC), activity, multisite protein phosphorylation and, 436
- PDI. *See* Protein disulfide isomerase (PDI)
- Pea
- cellular stress response, to salinity, 608*t*, 612
- defense response against pathogens, proteomics of, 588, 591*t*, 597
- proteome, changes in response to abiotic stresses, 621
- Peach trees, defense response against abiotic stress, transcriptomics/proteomics of, 622*t*, 623
- PEAKS, 42, 137
- Pectate lyase, pollen, as allergen, 271
- Pectinesterase, pollen, 267, 269
- Pectin matrix
- in cell wall, 293, 294*f*
- protein interactions with, 295
- Pectin methylesterase inhibitor(s) (PMEI)
- in cell wall, 301
- in pollen, 269–270
- PepNovo, 42, 137
- Peptidase, 92
- Peptide(s). *See also* Phosphopeptides
- amino acid sequence of, tandem MS analysis of, 36
- as biomarkers, 92
- from common protein precursor, 96
- definition of, 91
- de novo* sequencing, 40, 41*f*, 42, 135–138
- as diagnostics, 100, 101*t*
- differential display, 98–99
- as drugs (therapeutics), 100, 100*t*
- enrichment of, 43
- in human blood, 93
- identification of, 128–129
- mass profiling, using MALDI-MS, 26
- native, extraction from samples, 95
- proteases and, 101–102
- separation, 22, 146–147
- technology for, 93–95
- wheat endosperm, fragmentation of, 208–210
- Peptide fragmentation identification, 40–42, 41*f*
- Peptide hormone(s), precursors, 102

- Peptide mapping
 in situ, 14
 by SDS-PAGE, 14–15
- Peptide mass fingerprint (PMF), 40, 41 *f*
- PeptIdent, 595
- Peptide Prophet, 210
- Peptidomics, 91–104, 94 *f*
 advances in (future directions for), 103
 applications of, 100–101
 bioinformatics in, 97–98
 data mining in, 97–98
 2D CA-RP-LC-ESI-MS in, 100
 definition of, 91
 1D LC-MALDI in, 99–100
 key issues in, 95*b*
 mass spectrometry in, 95–97
 system/workflow in, 93
 technologies for, 93–95
 top-down sequencing in, 96
- Peptidyl vinyl sulfones (PVS), 49
- Perisperm, 191
- Peroxidase (POX)
 in cell wall, 300
 organ specificity of, 302, 303*t*
 expression of, in rice leaves, 246
 pollen, 267
 in root tissue, salt stress and, 227*t*, 234
 in rosette leaves, 301
- Peroxidase (POX) precursor, in root tissue,
 regulation of, 231*t*, 232
- Peroxidoxins, pollen, 267
- Peroxioredoxin, in root tissue, heavy metals and,
 227*t*, 235
- Peroxisomal and malate dehydrogenase
 (microtubule-binding protein), 280*t*
- Peroxisomal enzymes, 377–378
- Peroxisomal matrix proteins
 MT-binding activity, 283, 284, 285–286
 RNA-binding activity, 283, 285–286
- Peroxisomal multifunctional protein
 (microtubule-binding protein), 280*t*, 283
- Peroxisomal targeting signals (PTSs), 377–378
- Peroxisome, 377–390. *See also* Glyoxysomes
 cotyledonary, 378, 378 *f*
 functional classification of, 378, 378 *f*
 functions of, regulation of, analysis of, 387
 leaf, 378, 378 *f*
 metabolic profiling, 387
 organ-specific differentiation of, 378, 378 *f*
 proteins
 MALDI-TOF-MS analysis, 383
 PMF analysis, 383
 proteomics of
 of mammals, 379
 of plants, 379
 selection of species for, 379–380
 progress in, 386
 and transcriptomics, combination of, 387
 of yeast, 379
 root, 378, 378 *f*
 transcriptomics of, 378
 unspecialized, 378, 378 *f*
- PGIP. *See* Polygalacturonase inhibitor proteins
 (PGIP)
- PGK, in nucleus, 332, 334
- Pharmalyte, 17, 18, 18 *f*–19 *f*
- Phenol extraction, 39
- Phenylx, 130
 performance achieved, 134–135
 scoring function, 134
- Phloem, 239, 240 *f*
- Phosphatases
 in plasma membrane, 322
 suicide substrates, 49–51, 51*t*
 treatment of phosphorylated proteins/peptides,
 for detection of phosphorylation, 428
- Phosphate transporters
 in chloroplast, 344
 in plasma membrane, 321
- 6-Phosphogluconolactonase, anther, regulation of,
 256
- Phospholipase D (PLD), 282–283
- Phosphopeptides
 enrichment, 422–426
 by chemical derivatization, 424–426
 fragmentation techniques
 CID/MS3 as, 430
 electron capture dissociation as, 430
 electron transfer dissociation as, 430
 in mass spectrometry, 429–430
 PSD as, 430
- Phosphoproteins, 421–442
 enrichment, 422–426
 by antibodies and capture molecules, 423
 by chemical derivatization, 424–426
- Phosphoproteomics, 6
 advances in, 496
 of *N*-glycosylation of proteins in plants,
 455–468
 isotope-coded affinity analysis for, 473–475,
 474 *f*–475 *f*
 of leucine-rich repeat receptor-like kinases (LRR
 RLKs), 473–475, 474 *f*–475 *f*
 of plant STY kinases, 506
 of ubiquitin pathway, 443–454
- Phosphorylation, 6, 42–43, 485. *See also* Tyrosine
 phosphorylation
 amino acids in, 421–422

Phosphorylation, (*Continued*)

- analysis, in plant extracts, methods for, 493–494
 - analysis of, progress in, 421–422
 - of cell wall proteins, 302
 - detection of, 427–432
 - antibodies in, 427
 - dye technology and, 429
 - mutation analysis and, 429
 - phosphatase treatment in, 428
 - radioactive labeling for, 427–428
 - large-scale proteomics studies of, 437
 - of leaf proteins, 244
 - multisite, 421–422
 - large-scale proteomics studies of, 437
 - in plants
 - biological significance of, 434–437
 - and cell-to-cell trafficking, 435
 - and enzymatic activity, 436–437
 - and growth control, 436
 - and photosynthesis, 435
 - and protein–protein interactions, 434
 - and signal transduction, 434
 - and symbiosis, 435
 - studies of
 - advances in (future directions for), 438
 - progress in, 438
 - in plants, 502
 - pollen proteins mediating, 267
 - quantification of, 431–432
 - chemical derivatization strategies for, 432
 - ICP-MS–based methods for, 433, 433*f*
 - imaging methods for, 431–432
 - MS-based methods for, 432–434
 - standard peptides and, 433–434
 - and regulation of protein–protein association, 501–502
 - single-site, 421–422
 - studies of
 - advances in (future directions for), 438
 - progress in, 438
 - solid-phase screening method for, 489–492, 491*f*–492*f*
- Phosphorylation sites
- analysis, isotope-coded affinity approaches for, 473–475, 474*f*–475*f*
 - databases of, 428
 - determination of, 427–432
 - chemical derivatization strategies for, 427
 - Edman sequencing and, 428
 - identification of, by LC-MS/MS, 476
 - in vitro* analysis of, using peptide arrays, 492–493
 - prediction programs, 428
- PhosphoTyr-based signaling, 501–502

- PhosphoTyr phosphatases, plant, *in silico* identification of, 507–508
- Photoaffinity labeling, of kinases
 - in chemical proteomics, 55–57, 56*f*, 57*f*
 - with H-9 derivative, 55–56, 56*f*
- Photomultiplier tube (PMT), 112–114
- Photosynthesis, multisite protein phosphorylation and, 435
- Photosystem II (PS II), proteins, *C. reinhardtii*, phosphorylation sites in, 55
- Phragmoplast, 276
- Phytohormones, and germination, 192
- pI. *See* Isoelectric point (pI)
- Picea abies* L. Karst., cellular stress response, to heavy metals, 608*t*, 614
- Pinus strobus*, pollen tube, proteomics of, 272
- Pistil, 262
- Pisum sativum*. *See* Pea
- Plant(s), proteomics, 5
- Plant cell(s)
 - compartments in, 61–62
 - cultures, for studying protein localization, 73–74, 74*f*, 75*f*, 76*t*
 - organelles, 61–62
 - substructure of, 61–62
 - transport within, 62
- Plant membrane proteomics, 473
- Plant organelles, proteomics, 6
- Plant–pathogen interactions, secretory proteins in, 84–86, 88, 89
- PlantsP, 428
- Plant tissues, proteomics, 6
- Plasma, proteomics, 6
- Plasma membrane (PM), 309–326
 - characteristics of, 309
 - enzyme activities in, 309
 - as exchange barrier for ions and solutes, 309
 - functions of, 309
 - isolated sample, purity assessment, 311*t*–312*t*, 314–315
 - protein complexes, BN-PAGE of, 566
 - in protein exchange with environment, 309
 - proteins of
 - in cell signaling, 322–323, 322*f*
 - in cell-wall biosynthesis, 323
 - in cell-wall composition, 323
 - and cell wall proteins,
 - interactions/relationships of, 323
 - and cytoskeleton, 323
 - in defense, 323
 - in endocellular trafficking, 322*f*, 323
 - functional classification of, 321, 322*f*
 - functional survey of, 320–323, 322*f*
 - GPI-anchored, targeted extraction of, 317

- hydrophobic, 319–320, 320*f*
 - specific recovery of, 315–316, 316*f*
 - identification of, MS-based analysis for, 311*t*–312*t*, 318–319
 - lipid-anchored, 319–320, 320*f*
 - in lipid rafts (detergent-resistant microdomains), targeted extraction of, 317–318
 - in metabolism, 322*f*
 - phosphorylated, targeted extraction of, 318
 - in protein maturation–translation, 322*f*, 323
 - in protein modifications, 323
 - in protein turnover, 322*f*, 323
 - in proteolysis, 323
 - in secretion processes, 323
 - specific recovery of
 - alkaline treatment in, 315–316, 316*f*
 - C/M extraction in, 316, 316*f*
 - complementary strategies for, 315–316, 316*f*
 - salt treatment in, 315–316, 316*f*
 - synthesis of, 84
 - targeted extractions of, 317–318
 - in transport, 321–322, 322*f*
 - unknown, 322*f*, 323
 - proteomics of
 - advances in (future directions for), 324–325
 - challenges of, 310
 - complementary strategies for, 313
 - literature review, 310, 311*t*–312*t*
 - progress in, 311*t*–312*t*, 319, 323–324
 - purification of, 310–315
 - from cell suspensions *versus* whole plants or tissues, 310–313, 311*t*–312*t*
 - methods for, 313–314, 314*f*
 - two-phase partitioning method for, 311*t*–312*t*, 313–314, 314*f*
 - in signal transduction and perception, 309
 - Plasmodium falciparum*, protein composition, determination of, 67
 - Plastids, 351
 - protein complexes, BN-PAGE of, 566
 - protein targeting, 355–356, 357*f*
 - Plastoglobuli, chloroplast
 - metabolic activity of, 341, 345
 - proteomics of, 342*f*, 345
 - purification of, 343
 - PLD. *See* Phospholipase D (PLD)
 - PM. *See* Plasma membrane (PM)
 - PMC. *See* Pollen mother cells (PMC)
 - PMEI. *See* Pectin methylesterase inhibitor(s) (PMEI)
 - PMF. *See* Peptide mass fingerprint (PMF)
 - PMIs. *See* Protein–microbe interactions (PMIs)
 - Pollen
 - allergens, analysis of, 270–271
 - development of, 262
 - post-translational modifications in, 265
 - stages of, 249–250, 250*f*, 253, 262, 263*f*
 - proteome maps at, 253–254, 255*t*–256*t*
 - dispersal, 262
 - functions of, 262
 - germination, 267
 - maturation of, 261
 - mature
 - functional classification of, 265–268, 266*t*
 - proteins of
 - in cell cycle and DNA processing, 266*t*
 - in cellular transport and signaling, 266*t*, 267
 - in cell-wall remodeling, 266*t*, 267, 269, 269*f*, 271
 - in cytoskeletal dynamics, 266*t*, 267, 269, 269*f*
 - in defense, 266*t*, 267–268
 - in energy generation, 266, 266*t*
 - in metabolism, 266, 266*t*
 - in protein processing, 266–267, 266*t*
 - in stress response, 266*t*, 267–268
 - transcription-related, 266*t*
 - with unknown function, 266*t*, 268
 - reference maps, establishment of, 264–265
 - oleosins, 408
 - proteins specific to, 269–270
 - proteomics of, 261–274
 - advances in (future directions for), 272
 - progress in, 264, 271–272
 - transcriptomics of, 261–262
- Pollen grain, 250, 262, 263*f*
 - as biological model system, 262–264
 - collection/harvest of, 263–264
- Pollen mother cells (PMC), 250, 250*f*, 262, 263*f*
- Pollen tube, 262
 - growth, 266, 267
 - proteomics of, 272
- Polyacrylamide gel electrophoresis (PAGE), 11.
 - See also* Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE)
- Polygalacturonase, pollen, 267
- Polygalacturonase inhibitor proteins (PGIP), in cell wall, 301
 - post-translational modifications of, 302
- Polyubiquitin, pollen, 267
- Poplars. *See also* *Populus*
 - cellular stress response, to low temperature, 609*t*, 619–620
- Populus*, genome, sequencing of, 62
- Populus euphratica*, cellular stress response, to heat stress, 609*t*, 618

- Populus trichocarpa*, 5
- POR. *See* Protochlorophyllide oxidoreductase (POR)
- Pore gradient electrophoresis, 13–14
- Pore limit, 13–14
- Pore-limit electrophoresis, 13–14
- Post-translational modifications (PTMs), 2, 3–4, 5, 6, 22, 92. *See also* N-Glycosylation; Phosphorylation
- analysis of, computational search algorithms for, 43
- of cell wall proteins, 302
- diagnostic ions, 43, 43*t*
- identification of, 33–34, 41–42, 42–44
- labile, 43
- of leaf proteins, 244–245
- mass increments, 43, 43*t*
- mass spectrometry studies of, 153–159
- neutral losses, 43, 43*t*
- in pollen development, 265
- stable, 43
- Potassium deficiency, and differential expression proteomics in *Arabidopsis*, 155*t*
- Potato, 5
- Potato virus X (PVX), cell transformation methodology using, 71
- Is. See* Protein–protein interactions (PPIs)D
- PR-10a. *See* Probenazole-induced protein PBZ1 (PR-10a)
- Prevacuoles, 392
- Probenazole-induced protein PBZ1 (PR-10a), in root tissue, regulation of, 231*t*, 232
- Profilin(s)
- as allergens, 271
- anther, 254, 255*t*
- in plasma membrane, 323
- pollen, 267
- as pollen allergens, 271
- ProFound, 595
- Profound (search engine), 252
- Prokaryotes, proteome, complete description of, 2–3
- Prolamins, 196
- Proline knot, 410, 413
- Proline-rich proteins (PRP), analysis of, 297–299
- ProMEX, 428
- Proplastids, 351
- Protease(s), 92
- actions of, 101–102
- Arabidopsis*, characterization, 144*t*–145*t*, 160
- in cell wall, 300–301, 301*f*
- and peptides, 101–102
- in plasma membrane, 323
- suicide substrate, 51, 51*t*
- Protease inhibitors, 92, 101–102
- in cell wall, 301
- Proteasome
- 20S, anther, 252, 253, 255*t*, 256
- subunits, in plasma membrane, 323
- Protein(s)
- 14–3–3, 515–524
- functions of, 515
- physical properties of, 515–516
- in plants, 517–518
- advances in (future directions for), 522–523
- analysis, progress in, 522
- proteomics of, 518–522, 521*f*
- species distribution of, 515–516
- structure of, 516–517, 517*f*
- tissue distribution of, 515–516
- compartmentalization, in plant cells, 61–66
- covalent modifications of, 3
- cytoskeletal, 62
- in mature pollen, 266*t*, 267, 269, 269*f*
- enrichment, 47–48
- by inhibitor affinity chromatography, 52
- expression, proteomics of, 699–701
- filamentous, in plant cells, 62
- formation of, 3
- functional status of, proteomics of, 701–704
- functions of, 62
- determination of, 3–4
- and genes, matching, 2–3
- housekeeping, 85
- identification of, 128–129, 209–210
- with mass spectrometry, 40–44
- immunolocalization, 69–70
- isolation, 1
- maturation of, 3
- ”missing” *versus* ”uncharacterized,” 3
- molecular forms, 92
- plant
- classification of, 196
- solubility groups, 196
- precipitation of, 23, 39
- PTM-specific, enrichment of, 43
- purity, assessment of, 12
- rice leaf, protein extraction protocol for, 168–169, 168*f*
- secreted, 83–84
- signal peptide of, 85
- soluble, in *planta*- and *in vitro*-derived, in rice, 87, 88*f*
- structure of
- determination of, 671–692
- prediction
- ab initio* methods, 685–687, 688*f*

- computational approaches for, 678–687
 - by homology modeling, 678–682, 679*f*
 - by protein threading, 682–685, 683*f*
- subcellular localization, 704
 - experimental determination of, 62, 63, 66–69
 - and function, 62, 63
 - in vivo* imaging, 69–73
 - plant cell cultures for studying, 73–74
 - prediction, algorithms for, 63–66, 64*t*
 - performance comparison and outlook for, 66
 - transformation methodologies for studying, 71–72
- targeting sequences, 63
- three-dimensional (3D) structure of, 3
- Protein chemistry, 1
- Protein chip technology, 650*b*
 - in identification of calcium sensor targets, 653–655, 654*f*
- Protein complex(es), 557–559
 - functions of, 557
 - subunits, 558–559
 - tissue distribution of, 557–558
- Protein disulfide isomerase (PDI)
 - anther, 253, 255*t*
 - in wheat amyloplasts, 217
- Protein fragment complementation assays (PCA), 76–77, 118–119
- ProteinHarvester, 130
- Protein kinases
 - analysis, progress in, 486
 - assays, in cellular extracts, 488–489
 - substrates, 485–498
 - identification of
 - advances in (future directions for), 496
 - in vitro* analysis of phosphorylation sites
 - using peptide arrays and, 492–493
 - in vitro* strategies and methods for, 486–493
 - in vivo* strategies and methods for, 493–494
 - progress in, 486, 496
 - solid-phase phosphorylation screening for, 489–492, 491*f*–492*f*
- Protein–microbe interactions (PMIs), 587
 - biotrophic, 629–644. *See also* Symbioses
 - mutualist, 630
 - parasitic *versus* beneficial, 629–630
 - proteomics of
 - advances in (future directions for), 640–642
 - progress in, 640, 641*t*
 - endophytic parasitic, 629
 - gene expression in, 587–588
 - necrotrophic, 629
 - proteomics of, 588, 590*t*–591*t*
- Protein phosphatase 2A (PP2A), 283, 285
- Protein phosphorylation. *See* Phosphorylation
- Protein Prophet, 210
- Protein–protein interactions (PPIs), 2, 3–4, 557–559
 - analysis, TAP-tag technology for, 527–528. *See also* TAP-tagging technology
 - examination of, methods for, 67–69
 - in vitro* screening methods, 487–488
 - in vivo*, FRET study of, 74–77
 - mapping, 571–583
 - advances in (future directions for), 581
 - progress in, 580–581
 - Y2H system for, 574–578, 576*f*
 - multisite protein phosphorylation and, 434
 - specificity of, analysis of, 572–574
 - stable, 4
 - strength of, analysis of, 572–574
 - transient, 4
- Protein threading, 682–685, 683*f*
- Protein Tyr kinases (PTKs), 500
 - identification of
 - by comparative proteome bioinformatics, 505–506
 - pattern scanning in, 508–509
 - recall and specificity of, 509
 - proteomic complements, bioinformatic
 - identification of, 508–509
 - in tissue/organ development, 501
- Protein Tyr phosphatases (PTPs), 500–501
 - pseudophosphatase, 501
 - in tissue/organ development, 501
- Proteome(s), 2, 6
 - complete description of
 - in eukaryotes, 2–3
 - in prokaryotes, 2–3
 - definition of, 698*t*
 - dynamics of, 704–705
 - fractionation of, 39
 - proteins of, identification of, 2–3
 - quantitative analysis of, 34
- Proteomics, 166, 166*f*, 713–714. *See also*
 - Cell-mapping proteomics; Chemical proteomics; Expression proteomics; Structural proteomics
 - abundance-based, 715
 - activity-based, 48
 - applicability to all parts of biology, 5
 - applied, 5–6
 - and asynchronous flow of functional evidence
 - vs.* proteome-wide analysis, 505
 - bioinformatics of, 4–5
 - bottom-up *versus* top-down, 44
 - components of, 2, 3*t*
 - definition of, 2–5, 698*t*
 - in developing countries, 713–730

- Proteomics, (*Continued*)
- advances in (future directions for), 726
 - progress in, 725
 - development of, 1–2
 - differential quantitative, in 2D map format, 27–28
 - environmental, 153, 154*t*–157*t*
 - and evaluation of protein expression, 699–701
 - functional, 48
 - gel-based, 11–31, 167
 - advances in, 28–29
 - bioinformatics in, 107–125
 - future directions for, 29–30
 - methodology for, 112–120
 - strategy for, 112–120
 - goal of, 5, 6
 - human studies, 5–6
 - Internet resources for, 144*t*–145*t*
 - mammalian paradigms, 6
 - MS-based, 167
 - in plant biology, 166–167
 - plant studies, 5
 - advances in, reviews of, 179
 - practical, 5–6
 - research
 - in China, 719–720
 - in developing countries, 719–720
 - in India, 719, 720–721
 - targeted approaches, 159–160
 - top-down, 127
 - yeast studies, 5
- Proteomics data, presentation of, 706
- Protochlorophyllide oxidoreductase (POR), 351
- Protoplast(s), transformation with naked DNA, 71
- PRP. *See* Proline-rich proteins (PRP)
- PR proteins. *See* Pathogenesis-related protein(s) (PR proteins)
- Prunus persica*. *See* Peach trees
- PsbP domain proteins, in chloroplast lumen, 346
- Pseudomonas syringae*, and differential expression proteomics in *Arabidopsis*, 156*t*
- PTKs. *See* Protein Tyr kinases (PTKs)
- PTSS. *See* Peroxisomal targeting signals (PTSs)
- p*-value, 131
- PVS. *See* Peptidyl vinyl sulfones (PVS)
- PVX. *See* Potato virus X (PVX)
- QAE (quaternary amino ethyl) acrylamide, 20
- Q–Q–normal plot, 116, 117*f*
- QSTAR, in proteomics, 96
- Quadrupole (Q), 36–37, 44
- Quantile normalization, 118
- Quercus ilex* L. *See* Holm oak
- Radicle, 192
- RAN-binding protein, 334
- Rapeseed. *See* *Brassica napus*
- Rat, liver, secretome analysis in, 84
- Reactive group, 48*f*
- Reactive oxygen species (ROS), 377
 - metabolism, in roots, 234
- Receptor-like PTP (RPTP), 501
- Recognition unit, 48*f*
- Reflectron technology, 36, 37
- Regulatory proteins, in plasma membrane, 322
- Remorins, in plasma membrane, 322, 323
- Reporter ions, 96
- Reporter tag, 48, 48*f*
- Reversed-phase (RP) chromatography, for peptide extraction from sample, 95
- Reverse protein microarrays (RPMAs), for analysis of protein phosphorylation states in plant extracts, 494
- Reversibly glycosylated polypeptide, pollen, 267
- Rhizobial nitrogen-fixing (RNF) symbiosis, 630
 - and AM symbiosis, overlapping patterns of protein accumulation, 640, 641*t*
 - developmental features of, 631–633
 - early stages, proteomics of, 638–639
 - genomic resources and, 633–634
 - microsymbionts, proteomics of, 633–634
 - proteomics of, 630–634, 637–640
- Rhizobium*
 - plant cell transformation methodology using, 72
 - plant symbiosis, root proteins regulated by, 226*t*, 233
- Rhizoctonia solani*, rice defense response against, differential proteome profiles during, 590*t*, 596
- Ribosomal protein(s)
 - in MT proteome analysis, 284
 - pollen, 267
 - S12E, anther, regulation of, 256
- Rice, 5. *See also* *Oryza sativa* L.
 - anther
 - cold-responsive proteins, 254–256
 - development of, 249–250, 250*f*
 - proteome reference maps, 254
 - proteomics of, and transcriptomics, comparison of, 258
 - response to biotic and abiotic stress, 254–258
 - transcriptomics of, and proteomics, comparison of, 258
 - calcium sensors, identification of, bioinformatics approaches for, 648–649
- callus
 - differentiation of, 176
 - proteomics studies of, 176

- formation, plant hormones and, 232
- caropses, proteomics of, 195
- cellular stress response, to low temperature, 609*t*, 618–619
- chloroplast, proteomics of, 340–341
- cytoplasmic male sterility in, 257–258
- defense response against abiotic stress, transcriptomics/proteomics of, 622*t*, 623
- defense response against pathogens
 - differential proteome profiles during, 595–597
 - proteomics of, 588, 590*t*–591*t*
- developmental proteomics of, 172, 175*f*
- dynein, 281
- endophytic bacterial colonization, JA and, 234
- environmental proteomics of, 172, 175*f*
- genome, 527
 - sequencing of, 62
 - updated and enhanced data for, 166, 171–172
- heat stress and, proteomics studies of, 176–177
- kinase complexes, isolation of, TAP-tagging system for, 531*t*, 535–536, 536*f*
- leaf blades, etioplasts
 - intact, isolation of, 352, 353*f*
 - purification, 352, 353*f*
- leaf proteins, 245–247
 - expression of, during growth, 245–246
 - extraction protocol for, 168–169, 168*f*
 - extraction techniques for, 241
 - focusing and extraction in first and second dimensions, 169–170, 171*t*, 172*f*, 173*f*, 174*f*
 - salt stress and, 246–247
 - separation of, 245, 246*f*
 - silver nitrate staining, 170, 171*t*, 172*f*, 245, 246*f*
 - temperature and, 246
 - two-dimensional (2D) gel electrophoresis, 245, 246*f*
- male gametophyte development in, 254
- nuclear proteome, comparative analysis of, 335–336, 336*f*
- nucleus, proteomics of, 328
- plastid, protein targeting, 355–356, 357*f*
- pollen
 - mature
 - functional classification of, 265–268, 266*t*
 - reference maps, establishment of, 264–265
 - proteins of, isoforms, 265
 - proteomics of
 - advances in (future directions for), 272
 - progress in, 264, 271–272
 - proteomics studies, 174–175
- protein complex isolation from, TAP-tagging technology for, 527–542
- proteome, 165–178
 - changes in response to abiotic stresses, 620
 - reference map development, 169–170, 170*f*
- proteomics study(ies)/research, 165–166
 - accomplishments before 2006, 172–174
 - accomplishments in 2007, 174–177
 - advances in (future directions for), 177–178
 - bioinformatics and, 171–172
 - image analysis, 170–171, 172*f*–174*f*
 - methodology for, 167–172
 - progress in, 177
 - protein identification, 170–171, 172*f*–174*f*
 - simplified strategy for, 171, 174*f*
 - strategy for, 167–172
- putative protein Tyr kinases, 509–510
- root formation, plant hormones and, 232
- root tonoplast, plant hormones and, 395
- secretome analysis in, 84, 85–86
 - in planta*, 87, 88*f*
 - in vitro*, 87, 88*f*
- seedling, proteomics studies, 175–176
- spotted leaf mutant, proteomics studies of, 176
- vacuoles, proteomics of, 395
 - advances in (future directions for), 403–404
- Rice blast fungus, rice defense response against, differential proteome profiles during, 590*t*, 595–596, 601, 602*f*
- Rice yellow mottle virus (RYMV), rice defense response against, differential proteome profiles during, 590*t*, 597
- RLK proteins, in plasma membrane, 322
- RNA helicase, in MT proteome analysis, 284
- RNAi, with TAP-tagging technology, 536–537
- RNF. *See* Rhizobial nitrogen-fixing (RNF) symbiosis
- ROC curve, 131, 132*f*
- Root(s)
 - cell wall proteomics of, 296
 - development of, 223–224
 - proteomics of, 225–232
 - advances in (future directions for), 236
 - progress in, 236
 - dicot, 223, 224*f*
 - functions of, 223
 - interaction with environment, 223–224
 - lateral development of, proteomic studies of, 225
 - monocot, 223, 224*f*
 - mutants, study of, 225
 - proteins of, regulation of, 225–236, 226*t*–231*t*
 - proteome, 223–238
 - proteomics of, 6, 223–238
 - advances in (future directions for), 236
 - progress in, 236

- symbiosis with microorganisms, 223, 224*f*
- tissues of, 223, 224*f*
- Root apical meristem, 223, 224*f*
- Root hairs, 223, 224*f*
 - development of, proteomic studies of, 231–232
- Root nodules, proteomics of, 639–640
- ROSETTA toolkit, for *ab initio* protein structure prediction, 687, 688*f*
- Rotofor instrument, 19, 29
- RPMA*s*. *See* Reverse protein microarrays (RPMA*s*)
- rtcs* mutant, root protein regulation in, 225, 229*t*–230*t*
- RuBisCO, 147
 - in chloroplast, 345
 - expression of, in rice leaves, 246
- RuBisCO activase isoform precursor, *in planta*- and *in vitro*-derived, in rice, 87, 88*f*
- RYMV. *See* Rice yellow mottle virus (RYMV)
- Salicylic acid (SA)
 - and differential expression proteomics in *Arabidopsis*, 157*t*
 - and plant protein secretion, 84–85
- Salinity
 - cellular stress response to, proteomics of, 608*t*, 610–612
 - root proteins regulated by, 226*t*, 227*t*, 234–235
- Salt stress
 - and rice leaf proteome, 246–247
 - root proteins regulated by, 226*t*, 227*t*, 234–235
- Salt-stress-induced protein SalT, in root tissue, regulation of, 231*t*, 234
- Sample preparation
 - and identification of post-translational modifications, 43
 - for mass spectrometry, 39, 39*t*
 - for two-dimensional (2D) mapping, 23–25, 29–30
- SAR*s*. *See* Scaffold attachment regions (SAR*s*)
- Scaffold, 210
- Scaffold attachment regions (SAR*s*), 335
- ScanProsite, 300
- SCFG*s*. *See* Stochastic context-free grammars (SCFG*s*)
- Scoring function(*s*), 130–134
- SDS. *See* Sodium dodecyl sulfate (SDS)
- SDS-PAGE. *See* Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE)
- Search engines, 130, 133–135, 210, 252. *See also specific search engine*
- Secondary electron multipliers (SEM*s*), 38
- Sec proteins, in plasma membrane, 323
- Secretome, 83–90
 - analysis
 - advances in (future directions for), 89
 - methodology for, 86–87
 - in planta*, 86–87, 88*f*
 - state of the art, 87–89
 - strategy for, 86–87
 - in vitro*, 86–87, 88*f*
 - definition of, 83
- Secretory pathway(*s*)
 - in *Bacillus subtilis*, 84, 89
 - in plants, 83–84, 87, 88*f*, 89
- Seed. *See also* Cotyledons; Embryo; Endosperm; Perisperm
 - angiosperm, formation of, 191–192
 - development of, 191
 - metabolic control of, 203–204
 - proteomics of, 192–196
 - advances in (future directions for), 205
 - progress in, 204–205
 - endospermic, 192
 - germination, 191
 - hormonal control of, 192
 - metabolic control of, 203–204
 - protein levels during, regulation of, 201, 202*f*
 - proteomics of, 198–201
 - advances in (future directions for), 205
 - progress in, 204–205
 - mature, proteomics of, 196–197
 - nonendospermic, 192
 - oil bodies
 - proteins of, 407
 - proteome of, 410–412
 - oleosins, 408
 - proteins, post-translational modifications, during germination, 199–200
 - proteomics of, 6
 - teguments of, 191
- Seedlings
 - cultured in liquid medium, cell wall proteomics of, 296
 - etiolated, culture medium of, cell wall proteomics of, 302, 303*t*
 - post-germinative growth of, peroxisomes in, 377
- Seed plants, pollen tube, proteomics of, 272
- SEM*s*. *See* Secondary electron multipliers (SEM*s*)
- Senescence, leaf
 - as programmed cell death, 242
 - and proteomics, 242–244
- Sequest, 130, 210
 - performance achieved, 134–135
 - scoring function, 132–133
- Serine, phosphorylation, 43, 43*t*
- Serine carboxypeptidases, in cell wall, 301

- Serine hydrolase, irreversible inhibitor, as directed affinity-based probe, 49, 50*t*
- Serine protease, irreversible inhibitor, as directed affinity-based probe, 49, 50*t*
- Servalyt, 17, 18, 18*f*
- Shapiro–Wilk test, 116, 117*f*
- SH2 domains, phosphoTyr binding, 501–502
 in silico identification of, 508
- Sheath blight disease, rice defense response
 against, differential proteome profiles during, 596
- Short-chain dehydrogenase/reductase, in oil bodies, 408
- Signal peptide(s), 85
 analysis, 87
 secretory protein lacking, 85, 87, 88
- Signal perception, plasma membrane proteins
 involved in, 322–323
- SignalP software, 85, 87
- Signal transduction, 3, 4
 calcium-mediated, 645–658
 multisite protein phosphorylation and, 434
 plasma membrane proteins involved in, 322–323, 322*f*
 pollen proteins in, 266*t*, 267
 proteins
 in cell wall, 301, 301*f*
 in MT proteome analysis, 279*f*, 283, 284
- SILAC. *See* Stable isotope labeling technology (SILAC)
- Silver nitrate staining, of proteins in rice leaves, 170, 171*t*, 172*f*
- Silver staining, 25–26
 sensitivity in protein detection, 251, 252*f*
- Sinorhizobium*, plant cell transformation
 methodology using, 72
- Sinorhizobium meliloti*, plant symbiosis, root
 proteins regulated by, 226*t*, 233
- SKS5 multicopper oxidase homolog to SKU5, in *Arabidopsis thaliana* cell wall, 304
- Small Ub-related modifier. *See* SUMO
- SMART, 648
- SMRPs. *See* Symbiosis-related proteins (SMRPs)
- SNAREs, 640
 in plasma membrane, 323
- Sodium dodecyl sulfate (SDS), 11
- Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE), 12–17, 28
 advances in, 16
 buffers, conventional *versus* discontinuous, 13
 constant %*T* gels *versus* porosity gradients for, 13–14, 15*f*
 focusing, 16, 16*f*
 for leaf protein separation, 241
 linear gradient, 14, 15*f*
 peptide mapping by, 14–15
 porosity gradient, 14
 in two-dimensional (2D) mapping, 12
 uses of, 12–13
- Soft-ionization techniques, 33, 34
- Software
 for data integration, 129–130, 595
 2-DE analysis, 107–110, 108, 111, 118–119, 120–123
 variance introduced by, 115–116, 120–122, 121*f*
 imaging, 114
 for multiplexed analysis, 111
- Somatic embryogenesis, proteomics of, 201–203
- Sorghum, 5
- Sortins, 404
- Soybean. *See also Glycine max*
 cellular stress response, to herbicides, 608*t*, 616–617
 root hairs, proteomics of, 638
 symbiotic relationships, proteomic analysis of, 638–639
- Space–charge effect, 37
- Spectrum graph, 135–136, 136*f*
- Spinach
 chloroplast, proteomics of, 344
 plasma membrane protein complexes, BN-PAGE of, 566
- SPIRAL1 (microtubule-binding protein), 280*t*
- SPIRAL2/TOR1 (microtubule-binding protein), 280*t*
- Splicing factor-like protein, in root tissue, salt stress and, 227*t*, 234–235
- Spot detection
 automatic, algorithms for, evaluation of, 114–115, 115*f*, 120
 nonparametric, 108–109
 parametric, 108–109
 segmentation algorithms used in, 108
- Spot quantification, 109
- spr2/tortifolia1* mutant, 282
- SPS. *See* Sucrose-phosphate synthase (SPS)
- SRPGS. *See* SYPRO Ruby protein gel stain (SRPGS)
- Stable isotope labeling technology (SILAC), 153
 chemical, 153
 metabolic, 153
- Stains, compatible with mass spectrometry, 25–27
- Stamen, 249–250, 250*f*
- Starch granules, in wheat amyloplast, 215–216, 216*f*
- Stem(s), cell wall proteomics of, 296, 302, 302*t*, 303*t*

- Steroleosins, 411, 412–413
- Stochastic context-free grammars (SCFGs), 65
- Stomium, 250
- Stress response
 cellular, to abiotic stress, proteomics of, 605–628
 plant mitochondria and, 363–364
 pollen proteins in, 266*t*, 267–268
 proteomics of, 717–718
 subcellular proteome in, 599–601
- Stress tolerance, proteins related to, in root tissue, regulation of, 230*t*, 232
- Strictosidine synthase, in plasma membrane, 323
- Stroma, chloroplast, 339, 340*f*
 proteomics of, 342*f*, 345
- Stromules, 215, 216*f*
- Structural proteomics, 671–692
 advances in (future directions for), 687–689
 NMR spectroscopy in, 671, 675–677
 progress in, 671–672, 687
 X-ray crystallography in, 671, 672–675
- STY kinases, plant, genome-wide and Tyr phosphoproteome analysis of, 506
- Suaeda aegyptiaca*, cellular stress response, to salinity, 608*t*, 611–612
- SUBA database, 152
- Subcellular compartments
 protein localization in, 66–67. *See also* Protein(s), subcellular localization
 proteomics of, 144*t*–145*t*, 146*f*, 148–150, 149*f*, 328
- Subtilisins, in cell wall, 301
 post-translational modifications of, 302
- Succinyl-CoA synthetase, in root tissue, heavy metals and, 227*t*, 235
- Sucrose-phosphate synthase (SPS), activity, multisite protein phosphorylation and, 436–437
- Sucrose synthase, in MT proteome analysis, 284
- Sugar beet, cellular stress response, to drought, 608*t*, 609–610
- Suicide substrates, 48, 49, 51, 51*t*
- Sulfate adenyltransferase 4, expression in root development, regulation of, 225, 230*t*
- SUMO, 444
 analysis of, 451
- Superoxide dismutase
 expression of, in rice leaves, 246
 in root tissue
 heavy metals and, 227*t*, 235
 regulation of, 230*t*, 232
- Support vector machine (SVM), in prediction of plant protein localization, 64*t*, 65
- Surface-enhanced laser desorption/ionization (SELDI), 95
- Surface-enhanced laser desorption/ionization (SELDI)–TOF–MS, 92
- SVM. *See* Support vector machine (SVM)
- Symbioses. *See also* Rhizobial nitrogen-fixing (RNF) symbiosis
 AM, 630
 appressorium stage, proteomics at, 636–637
 developmental features of, 631–633
 proteomics of, 630–634
 recent progress in, 634–637
 and RNF symbiosis, overlapping patterns of protein accumulation, 640, 641*t*
 genomic resources and, 633–634
 legume and microorganisms, proteomic studies of, 182–183, 184, 187, 226*t*, 233
 multisite protein phosphorylation and, 435
 plant/bacterial, proteomic studies of, 182–183, 184, 226*t*, 233, 629–644
 plant/fungal, proteomic studies of, 182–183, 226*t*, 233, 629–644
 plant/microbe
 proteomics of, 629–644
 and root protein regulation, 226*t*, 228*t*, 233–234
 root and microbes in, 223, 224*f*, 226*t*, 233–234
 species specificity of, 234
- Symbiosis-related proteins (SMRPs), 630–631
- Symbiosomes, 631, 633
- Synechocystis*, cellular stress response, to salinity, 608*t*, 612
- Syntaxins, 640
 in plasma membrane, 323
- SYPRO Ruby, 26, 113, 113*f*, 114, 118, 123
 sensitivity in protein detection, 251, 252*f*
- SYPRO Ruby protein gel stain (SRPGS), 26
- Systems biology, 4, 159–160, 161
 comprehensiveness required for, 696–698
 data collection intervals and, 699
 definition of, 698*t*
 dynamics information required for, 699
 information about species required for, 698
 modeling approaches, 707–708
 proteomics and, 695–712, 697*f*
- TAC. *See* Tubulin-affinity chromatography (TAC)
- TAN1 (microtubule-binding protein), 280*t*, 281
- Tandem affinity purification (TAP) tag, 68
- Tapetum, anther, 249–250, 250*f*, 262, 263*f*
- TAP-tagging technology, for protein complex isolation
 advances in (future directions for), 541, 555
 advantages of, 538, 543

- applications of, 527–528, 543
 - from *Arabidopsis*, 543–556
 - results, 550*f*, 553–554
 - from budding yeast, 544, 545*f*
 - in calcium signaling networks, 655
 - crude extract for, preparation of, 532
 - fused target protein and associated complex
 - generation of, 546–548
 - purification of, 533–535
 - major concerns with, 538–539
 - methodology for, 528, 529*f*
 - modifications/alterations to, 536–537, 544
 - pEarlyGate, 546*t*
 - and pre-purification cross-linking step, 537
 - progress in, 540–541, 543–546, 554–555
 - protein complex purity, improving, 552–553
 - protein complex recovery, improving, 552–553
 - in rice, 527–542
 - study of kinase complexes, 531*t*, 535–536, 536*f*
 - scaling up, 552–553
 - and single-step affinity purification, 537
 - sTAPa, 546, 546*t*
 - systems available for (versions of), 528, 544, 546*t*
 - TAPa, 544–546, 545*f*, 546*t*, 548, 554–555
 - TAPi, 544, 546*t*, 548
 - TAP-tag for, 543–544
 - structure of, 528–529, 530*f*
 - transgenic plant material for, 530–532, 531*t*
 - functional characterization of, 548–549
 - preparation of, 549–551, 550*f*
 - trouble-shooting for, 539–540
 - using mutants, 536–537
 - using RNAi, 536–537
- TargetP, 346
- TCA. *See* Trichloroacetic acid (TCA)
- Technology, advances in, 1, 11–12
- Temperature. *See also* Heat stress
 - and anther proteome, 254
 - and differential expression proteomics in *Arabidopsis*, 154*t*
 - low, cellular stress response to, proteomics of, 609*t*, 618–619
 - and rice leaf proteome, 246
 - and wheat grain proteome, 213–214
- TFs, in nucleus, 334
- Thioredoxin, in seeds, 199
- Threonine, phosphorylation, 43, 43*t*
- Thylakoid membrane, 339, 340*f*
 - formation of, 351–352
 - functions of, 345
 - proteomics of, 341, 345–346
 - purification of, 342*f*, 343
- TIGR
 - Rice Genome Annotation project, 171
 - rice genome studies, 166, 171
 - rice resource database, 171
- Time of flight (TOF) mass analyzer, 37, 44
- Titanium, in metal oxide affinity chromatography, for phosphopeptide and phosphoprotein enrichment, 425
- TMV. *See* Tobacco mosaic virus (TMV)
- Tobacco
 - BY2 cellular system, for studying protein
 - localization, 73, 74*f*
 - cellular stress response
 - to nutrient availability, 608*t*, 613–614
 - to salinity, 608*t*, 612
 - plasma membrane proteomics, 312*t*, 313
 - proteome, changes in response to abiotic stresses, 620–621
- Tobacco mosaic virus (TMV)
 - cell transformation methodology using, 71
 - spread of, multisite protein phosphorylation and, 435
- Tocopherol, synthesis of, 345
- alpha-Tocopherol cyclase VTE1, in chloroplast, 345
- Tomato, 5
- Total spot volume (TSV) normalization, 111
- Total valid spot volume (VSV) normalization, 111
- TPS. *See* Trehalose-6-phosphate-synthase (TPS)
- TPT. *See* Triosephosphate translocator (TPT)
- Transcriptomics, 166, 166*f*, 714–716
 - Arabidopsis* pollen, 261–263
 - and proteomics, comparison of, 268–270, 269*f*
 - peroxisome, 378
 - pollen, 261–262
 - and proteomics, comparison of, 305
 - rice anther, and proteomics, comparison of, 258
- Transgenic plants, proteomics of, 717–719
- Translation
 - autoregulation, microtubule-binding proteins and, 285–286, 286*f*
 - nuclear proteins involved in, 334*f*, 335
 - proteins involved in, in MT proteome analysis, 279*f*, 283
 - significance of, 285
- Translational initiation factor
 - in MT proteome analysis, 283
 - pollen, 267
- Translationally controlled tumor protein, anther, 252, 255*t*
- Translation inhibitor(s), pollen, 267
- Translocon complex in inner envelope membrane (Tic), 355–356

- Translocon complex in outer envelope membrane (Toc), 355–356
- Transport proteins
in chloroplast, 344
in plasma membrane, 321–322, 322*f*
- Trehalose-6-phosphate phosphatase, in root tissue, heavy metals and, 227*t*, 235
- Trehalose-6-phosphate-synthase (TPS), activity, multisite protein phosphorylation and, 437
- Triacylglycerol hydrolase, in oil bodies, 408
- Trichloroacetic acid (TCA), 23
- Trichloroacetic acid (TCA)/acetone extraction buffer (TCAAEB), 168–169, 168*f*
composition, 168–169, 169*f*
- Trichloroacetic acid (TCA)/acetone precipitation, 23, 39
- Triosephosphate translocator (TPT), in etioplast, 354–355
- Triple quadrupoles (triple quad/Q), 37
- True negatives (TN), 130–131
- Trypsin
in sample preparation, 39, 39*t*
in vacuolar proteomics, 399
- Tubulin
 α -, 276
 β -, 276
pollen, 267, 270
- Tubulin-affinity chromatography (TAC)
in large-scale MT proteome studies, 283
in purification of microtubule-binding proteins, 278–280, 279*f*, 282–283
- Twin-arginine translocation (TAT) system, and protein import into thylakoid membrane, 346
- Two-dimensional (2D) gel electrophoresis (2-DGE), 1, 22, 147
advances in (future directions for), 29–30, 123–124
analysis, tools for, 108
background adjustment in, 109
cell fractionation and, 67
computational analysis of, 108
data
distribution pattern, 116–118, 117*f*, 122
variance analysis, 116–118, 117*f*, 122
image acquisition in, 112–114, 113*f*
image warping in, 109–110, 110*f*
for leaf protein separation, 241
methodology for, 112–120
normalization methods for, 110–112
post-electrophoretic variance in, 107
spot detection in, 108–109
spot matching in, 109
spot quantification in, 108–109
statistical analysis, 118–119
strategy for, 112–120
of wheat endosperm proteins, 210
- Two-dimensional (2D) mapping, 22–28, 27–28, 28
advances in (future directions for), 29–30
development of, 12, 13
differential processing, 29–30
DNA removal for, 23, 24*f*
interfering substances, 23
isoelectric focusing (IEF) in, 23
maximum resolution in, 14
nucleic acids and, 23
sample preparation, 23–25
advances in (future directions for), 29–30
alkylation step, 23–25
trichloroacetic acid (TCA)/acetone precipitation step, 23
SDS-PAGE for, 14, 15
spot overlap in, 29–30
- Tyrosine, phosphorylation, 43, 43*t*
- Tyrosine phosphatases, irreversible inhibitor, as directed affinity-based probe, 49, 50*t*
- Tyrosine phosphorylation
in eukaryotes
mechanism of, 499
relevance of, 499
pathways, enzymes involved in, 500–501
in plants, 499–514
developmental conditions and, 504
early evidence on, 502
in light-regulated signaling, 502–503
and regulation of plant bending, 503
and regulation of stomatal opening, 503
reversible, and response to phytohormones, 504–505
studies of
advances in (future directions for), 511
progress in, 510–511
in tissue/organ development, 501
- Ub. *See* Ubiquitin (Ub)
- UBDs. *See* Ubiquitin (Ub)-binding domains (UBPs)
- Ubiquitin (Ub)
functions of, 443
protein modification by, 443–444
- Ubiquitin (Ub)-activating enzyme (E1), 443–444, 444*f*
- Ubiquitinated proteins
analysis of
advances in (future directions for), 452
methods for, 445–448, 445*f*
MS technique for, 448–449
progress in, 444–445, 451–452
results, 448–451

- strategies for, 445–448, 445*f*
- by trypsin digestion and MS, 446–447, 447*f*
- polyUb chain, structure of, analysis, 444*f*, 447*f*, 449–450
- Ubiquitination, 43
- Ubiquitin (Ub)-binding domains (UBPs), 444
- Ubiquitin (Ub)-binding proteins (UBPs), 444
 - analysis of, 450–451
- Ubiquitin (Ub)-conjugating enzyme (E2), 443–444, 444*f*
- Ubiquitin (Ub) ligase (E3), 444, 444*f*
 - analysis of, 450
- Ubiquitin (Ub)-like proteins (ULPs), 444
 - analysis of, 451
- Ubiquitin (Ub) pathway, 443–454
 - DUBs, analysis of, 450–451
 - proteome analysis of, progress in, 444–445
 - regulatory stages in, 444, 444*f*
- UBPs. *See* Ubiquitin (Ub)-binding proteins (UBPs)
- UDP-glucose pyrophosphorylase, anther, regulation of, 257
- UDP-Glucuronic acid decarboxylase, anther, regulation of, 257
- ULPs. *See* Ubiquitin (Ub)-like proteins (ULPs)
- UniProt database, 129
- UniprotKB/Swiss-Prot database, 129
- V8 (endoprotease), in vacuolar proteomics, 399
- Vacuolar processing enzyme (VPE), 197
- Vacuolar sorting receptors (VSRs), 392
- Vacuoles
 - of animals, 391
 - of plants, 391–406
 - central, 391–392
 - enzymes of, 392
 - functions of, 391–392
 - genetic studies of, 392
 - isolation of, 395–396
 - high-resolution MS and, 396
 - LC-MS/MS *versus* MALDI-TOF-MS for, 396
 - lytic, 392
 - multiple types of, 391–392
 - protein delivery to, 392
 - protein storage, 391–392
 - proteome, defining, LOPIT strategy for, 396–397
 - proteomics of, 392–395
 - advances in (future directions for), 403–404
 - differential fractionation in, 396
 - differential proteolysis in, 399
 - differential scanning in, 398–399
 - gel-based, *versus* MudPIT, 396–397
 - label-free quantitative approach, 400–402, 400*f*
 - MS-based quantitative approaches, 399–401
 - progress in, 402–403
 - sequence coverage of peptides and proteins in, 396–399
 - SIL-assisted MS in, 399–400
 - size of, 391
 - of yeast, 391
- V-ATPase, 639
- Vigna unguiculata*. *See* Cowpea
- 2-Vinyl pyridine (2-VP), alkylation with, 24–25, 25*f*
- Vinyl sulfones, 49, 50*t*
- Virus
 - cell transformation methodology using, 71
 - rice defense response against, differential proteome profiles during, 590*t*, 597
- Vitis vinifera*. *See* Grapevines
- VPE. *See* Vacuolar processing enzyme (VPE)
- VSRs. *See* Vacuolar sorting receptors (VSRs)
- WAK. *See* Wall-associated kinases (WAK)
- Wall-associated kinases (WAK), 508
- Wheat
 - allergens, identification of, 215
 - amyloplast
 - amino acid synthesis in, 217–219, 218*f*
 - branched-chain amino acid pathway in, 217–219, 218*f*
 - functions of, 215
 - isolation of, 216–217
 - metabolic pathways, 217–219, 218*f*
 - proteins of, identification of, 217
 - proteome, 215–219
 - proteomics of
 - advances in (future directions for), 219–220
 - progress in, 219
 - starch granules in, 215–216, 216*f*
 - cellular stress response
 - to heat stress, 609*t*, 618
 - to herbicides, 608*t*, 616
 - to nutrient availability, 608*t*, 613
 - cultivars, identification of, proteomics and, 214
 - defense response against pathogens, proteomics of, 588, 591*t*, 597–598
 - endosperm
 - proteins from, 193*f*
 - proteome, 207–215
 - proteomics of, 196
 - advances in (future directions for), 219–220
 - progress in, 219
 - flour, protein fractions, characterization of, 214–215
 - genome of, 208
 - grain

- Wheat (*Continued*)
 amyloplast proteome, 215–219
 endosperm proteome, 207–215
 proteomic studies of, 211*t*–212*t*
 interactome walking strategy for, 580
 proteins, gene families encoding, 208
 proteomic studies of, 211*t*–212*t*
 advantages and disadvantages of, 208
 progress in, 208
- Wheat germ
 albumin proteins from, 193*f*
 cell-free protein synthesis system
 application to functional proteomics,
 663–665, 664*t*, 665*t*, 666*f*
 application to structural proteomics, 665–667
 development of, 662–663
 procedures for, 662–663, 663*f*
- White spruce, somatic embryogenesis in,
 proteomics of, 203
- Xanthomonas oryzae* pv. *oryzae*, rice defense
 response against, differential proteome
 profiles during, 590*t*, 596–597
- Xoo*. *See Xanthomonas oryzae* pv. *oryzae*
- X-ray crystallography
 and NMR spectroscopy, for protein structure
 determination, comparison of, 678
 protein structure determination by, 671,
 672–675
- X!Tandem, 130, 210
 performance achieved, 135
- XYL1 alpha-xylosidase, in *Arabidopsis thaliana*
 cell wall, 304
- Xylem, 239, 240*f*
 differentiation of, proteins related to, 296
- Xyloglucan-specific endoglucanase inhibitor
 protein (XEGIP), tomato, homolog to, in
 Arabidopsis thaliana cell wall, 304
- Yariv-affinity chromatography, of arabinogalactan
 proteins, 296, 297
- Yeast. *See also* Y2H system
 interactome, 574
 protein composition, determination of, 67
 protein interaction mapping for, 577
 proteomics, 5
 two-hybrid (Y2H) analyses, 2
- Yeast fluorescent protein (YFP), 75
- YFP. *See* Yeast fluorescent protein (YFP)
- Y2H system
 for identification of calcium sensor targets,
 652–653
 for protein–protein interaction mapping,
 574–578, 576*f*
- ZE. *See* Zone electrophoresis (ZE)
- Zea mays*. *See* Maize
- Zinc
 and root formation in rice, 232
 root proteins regulated by, 230*t*, 232
- Zinc-binding proteins, in root tissue, 235
- Zirconia, in metal oxide affinity chromatography,
 for phosphopeptide and phosphoprotein
 enrichment, 425
- Zone electrophoresis (ZE), 11–12. *See also*
 Capillary ZE-MS (CZE-MS)
- z-score, 131
- Zwischel gene mutants, 282

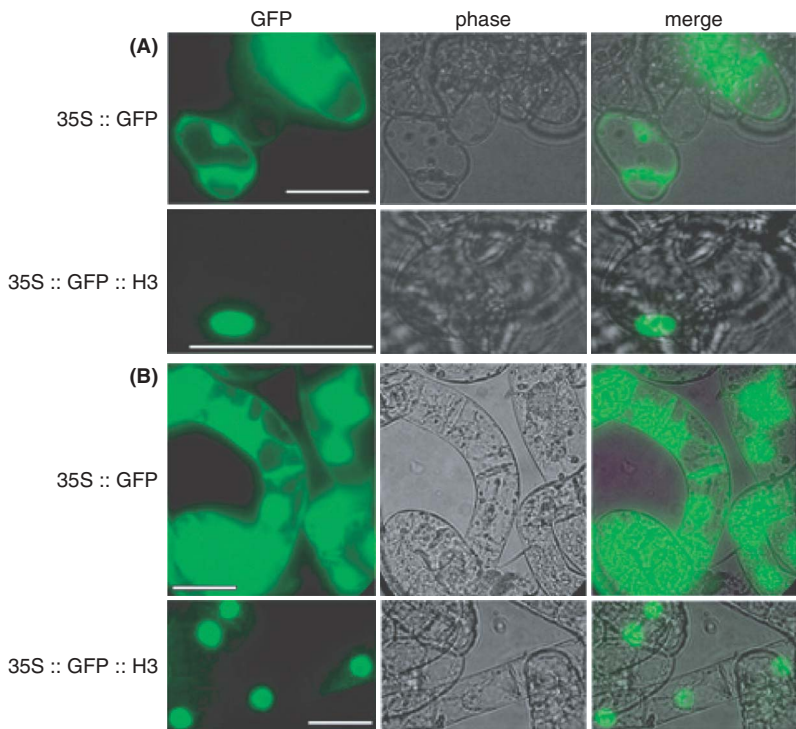


FIGURE 5.1 Comparison of GFP fusion protein localization in *Arabidopsis* culture cells (A) and tobacco BY2 cells (B). *Arabidopsis* cells have been transiently transformed, while BY2 cells were transformed and selected to produce stable cell lines. Patterns for GFP (unfused) and GFP-histone are shown. Bar represents 20 μm . See page 74 for text discussion of this figure.

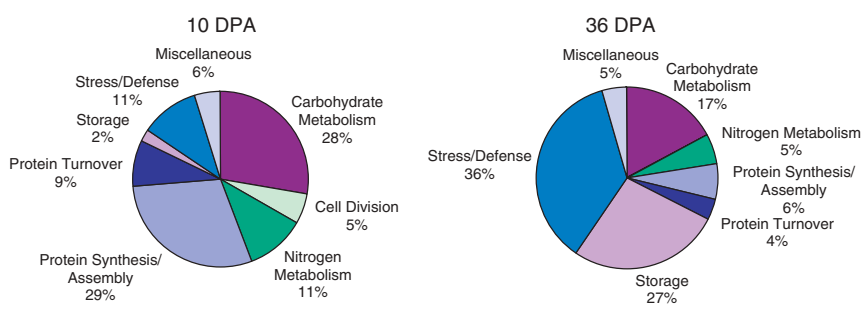


FIGURE 14.2 Functional distribution of proteins recovered in the albumin and globulin fraction of wheat endosperm isolated from developing grain harvested 10 and 36 days post-anthesis. The proteins comprising the functional categories can be found in Vensel et al. [4]. The globulins are the principal storage proteins in the albumin/globulin fraction. The miscellaneous category contains proteins that function in ATP interconversion, lipid metabolism, signal transduction, and transport. See page 213 for text discussion of this figure.

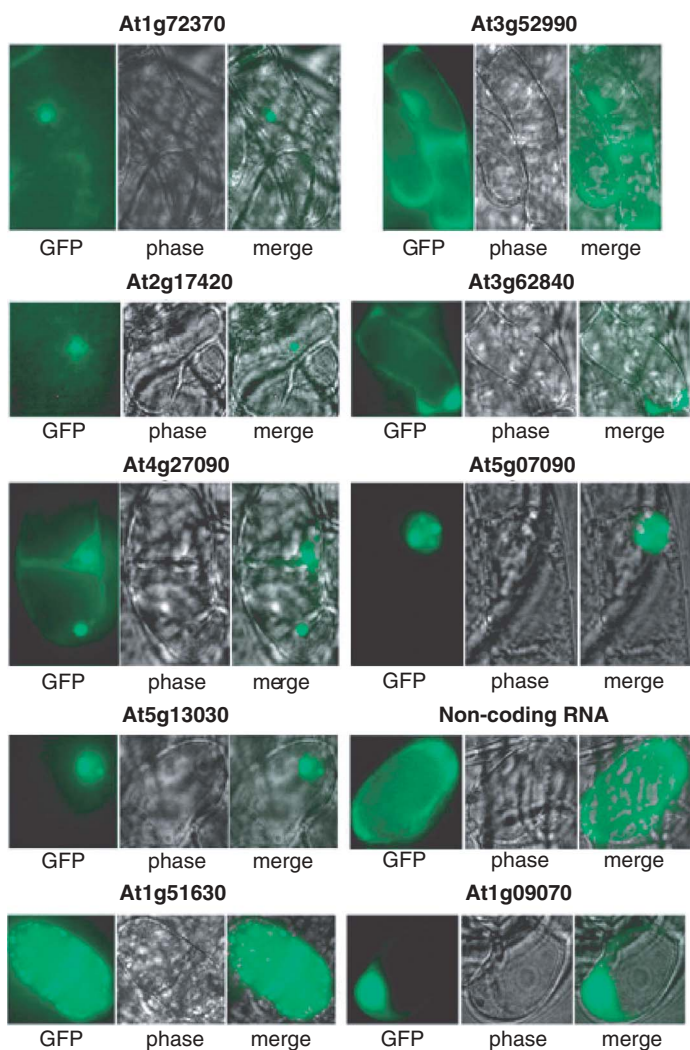


FIGURE 5.2 GFP-fusion localization showing a selection of patterns from a GFP localization screen based on transient transformation of *Arabidopsis* cells. Gene names are shown below the images and described in Table 2. (Bar represents 10 μm , except in At5g07090 where bar represents 7 μm). See page 75 for text discussion of this figure.

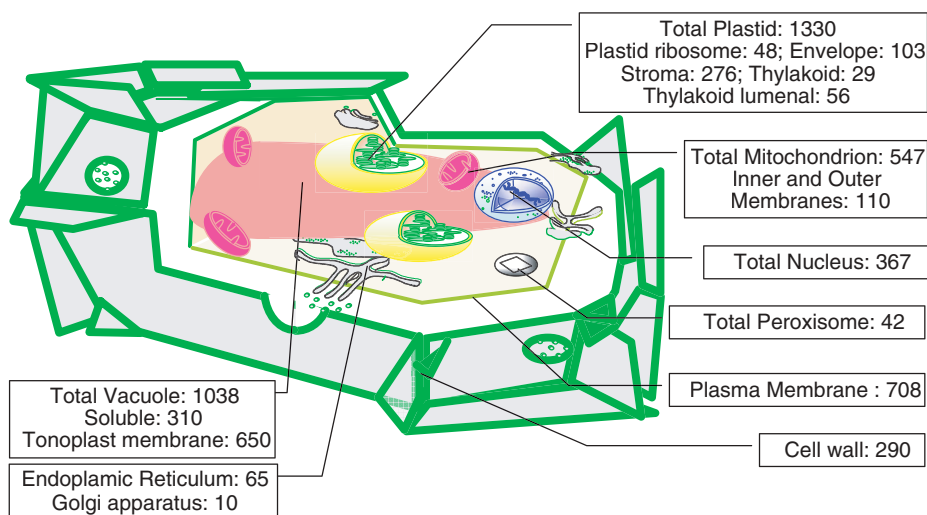


FIGURE 10.2 A. *thaliana* proteins identified by proteomics and their repartition in subcellular compartments (December 2006). See page 149 for text discussion of this figure.

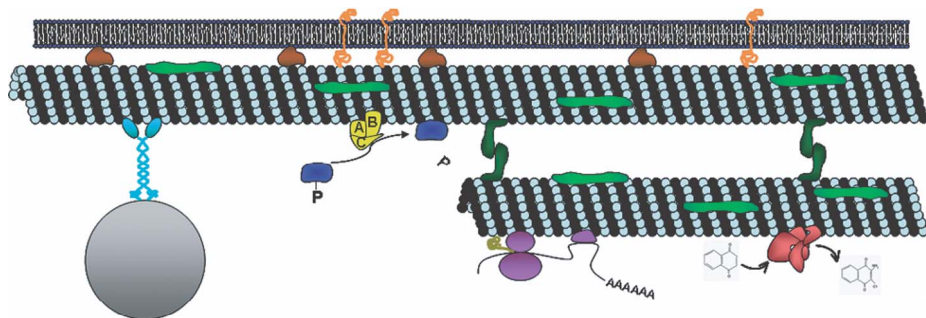


FIGURE 19.2 A model depicting possible interactions between MTs and MT-binding proteins. Structural MAPs are important in stabilizing MTs (light green), mediating the bundling of MTs (dark green), and anchoring MTs to the PM (brown). Motor proteins (turquoise) transport their cargo along MTs. Communication between MTs and the cell wall across the PM may involve TM proteins (orange). Metabolic complexes (red) may assemble on MTs to facilitate the flux of material through a metabolic pathway. Evidence for mRNA-binding protein and ribosome (violet) interactions with MTs may be important in mRNA localization and anchoring, as well as translational control. Signaling events involving protein phosphatases and kinases (yellow) could modify downstream signaling components or regulate the binding of proteins to MTs (blue). Other MT-binding proteins, such as EB1, are well-characterized in plants but are not depicted in this figure. See page 286 for text discussion of this figure.

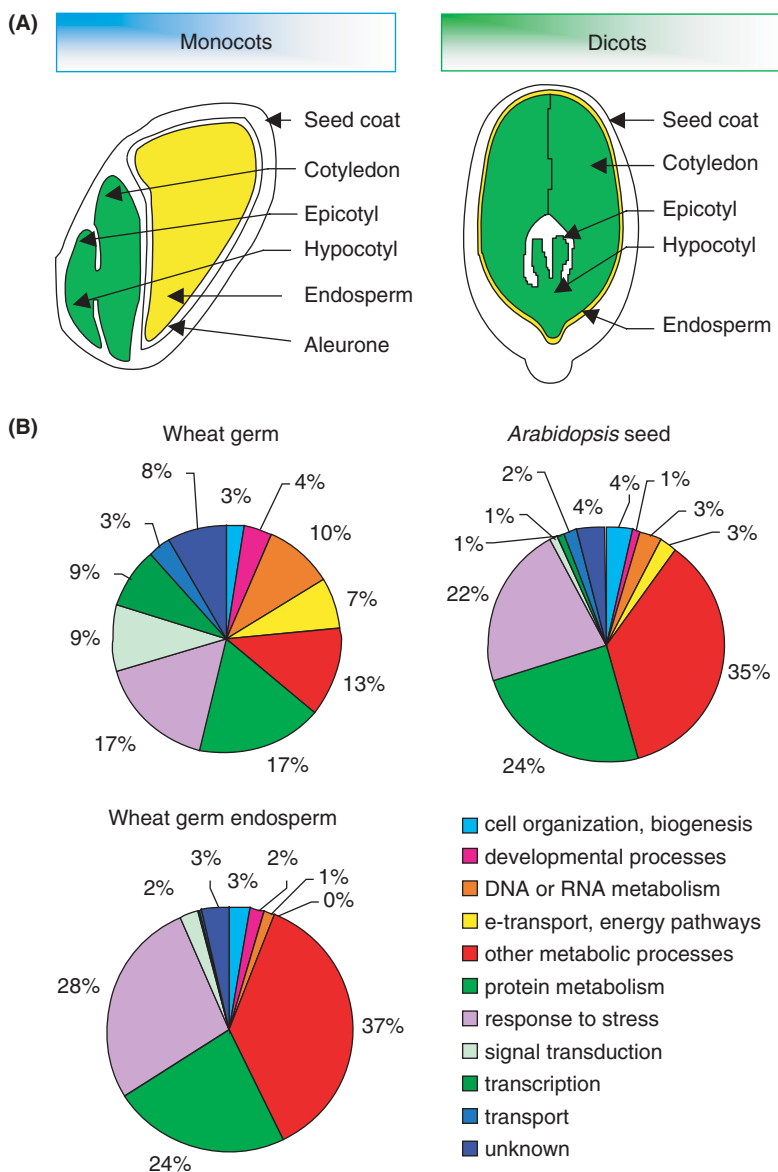


FIGURE 13.1 Structure of seeds and their proteomes. **(A)** Seed structure of monocots and dicots. **(B)** Albumin proteins from wheat germ (352 proteins analyzed; reference [13]), mature *Arabidopsis* seeds (358 proteins analyzed; reference [20]), and wheat endosperm (198 proteins analyzed; reference [7]). See page 193 for text discussion of this figure.

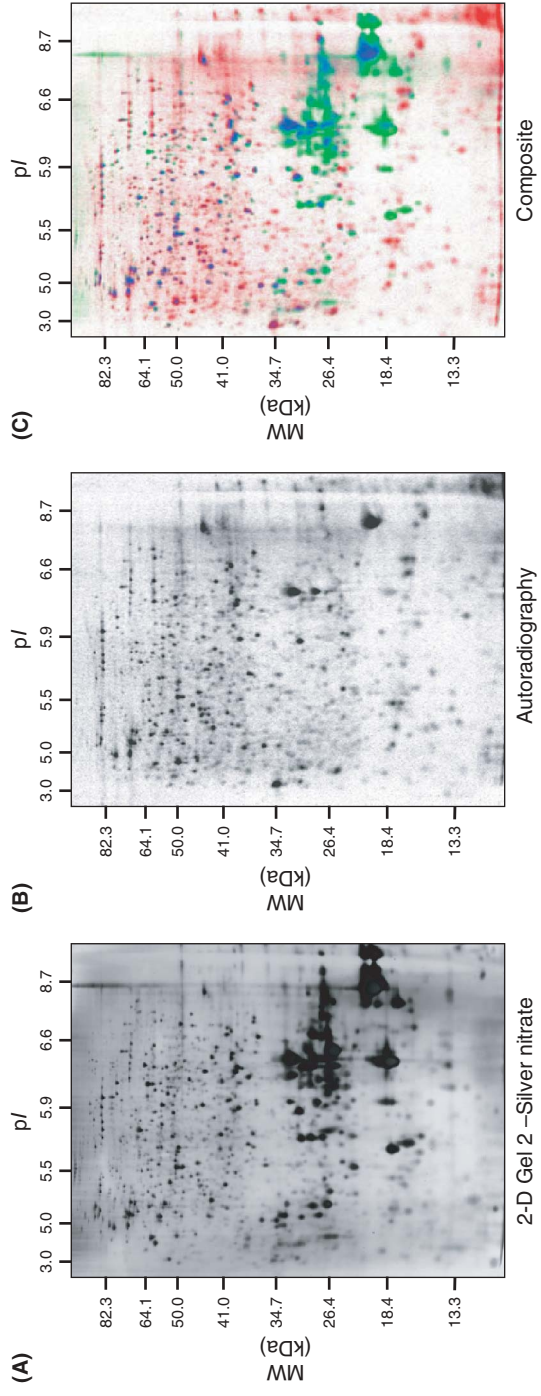


FIGURE 13.3 *De novo* protein synthesis during germination of *Arabidopsis* seeds (from Job et al., unpublished results). Seeds were incubated for 24 h in the presence of [35 S]-Met. Proteins were extracted and then submitted to 2-DGE, and the radiolabeled proteins were revealed by phosphorimager analysis. **(A)** Silver-stained gel. **(B)** Radiolabeled proteins from same gel as in A. **(C)** Superimposition of 2D patterns shown in A and B using false colors. Spots in red, proteins detected only by autoradiography; spots in green, proteins detected only by silver staining; spots in blue, proteins detected both by silver staining and autoradiography. See page 202 for text discussion of this figure.

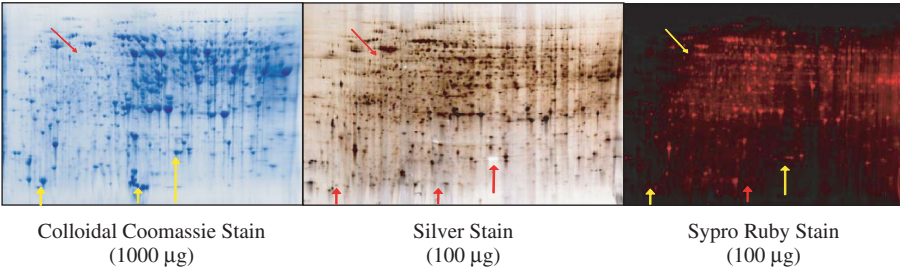


FIGURE 17.2 Comparison between colloidal CBB staining, silver staining, and SYPRO Ruby staining. Red arrows indicate protein spots that are not detected or negatively stained, and yellow arrows indicate properly stained protein spots. See page 252 for text discussion of this figure.

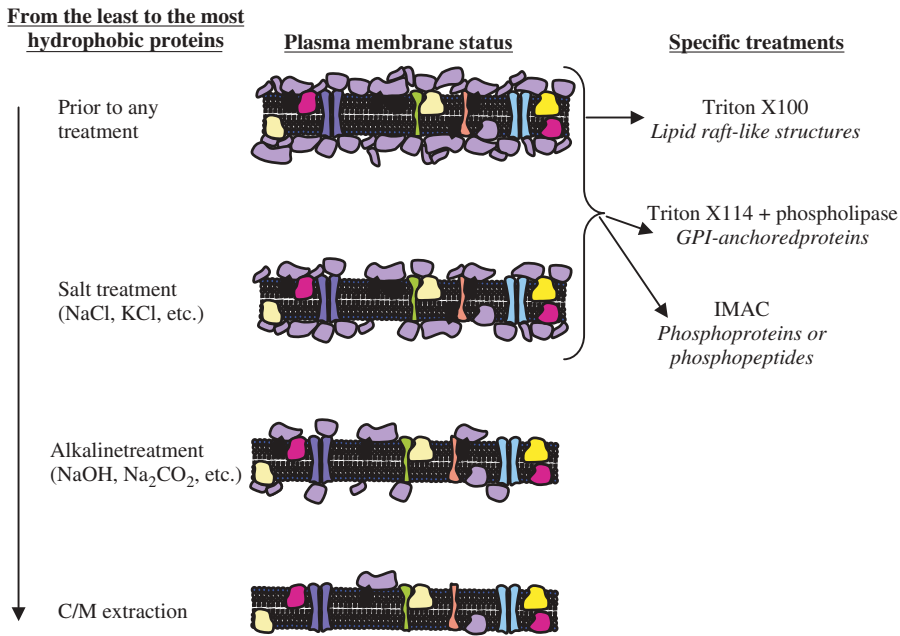


FIGURE 21.2 Outline of complementary strategies for the specific recovery of PM proteins. The left part of the figure indicates strategies aiming at the recovery of the largest proteome. They allow isolating proteomes differing at least partly by the physicochemical properties of the identified proteins. The compilation of these proteomes led to the more extensive overview of the PM proteome so far. On the right part of the figure, different strategies devoted to the retrieve specific protein categories are listed. See page 316 for text discussion of this figure.

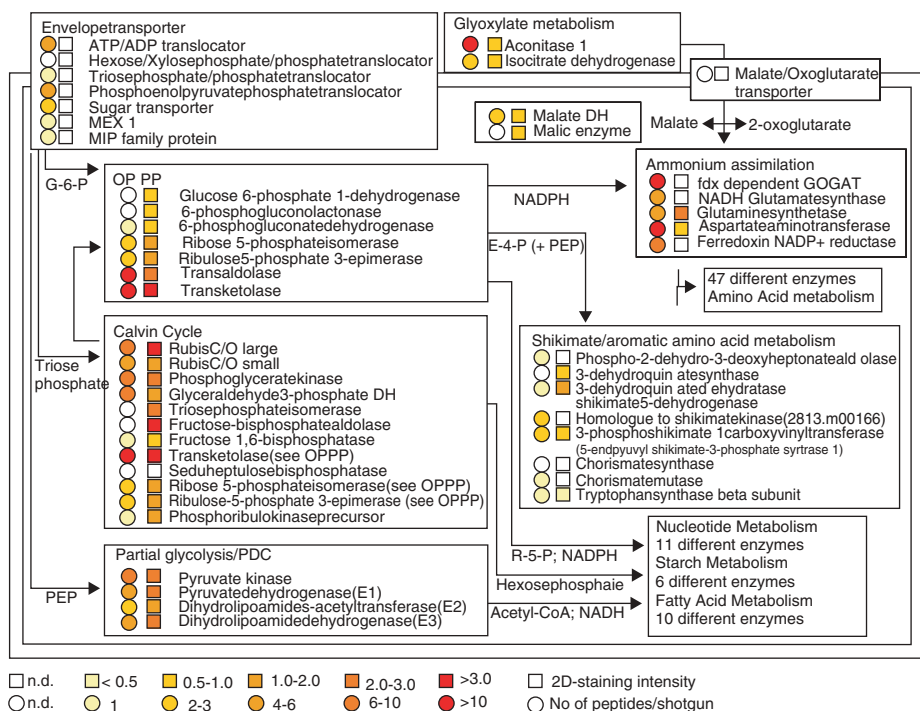


FIGURE 24.2 Color-coded schematic illustration of characteristic metabolic pathways in heterotrophic plastids. Missing protein identifications from known plastid pathways are depicted by white boxes and circles, and abundances of identified proteins are provided as the number of peptides in a shotgun proteome analysis (circles) and the staining intensity in the 2D-map (squares). E-4-P, erythrose-4-phosphate; G-6-P, glucose-6-phosphate; OPPP, oxidative pentose phosphate pathway; PDC, pyruvate dehydrogenase complex; PEP, phosphoenole pyruvate; R-5-P, ribulose-5-phosphate. See page 354 for text discussion of this figure.

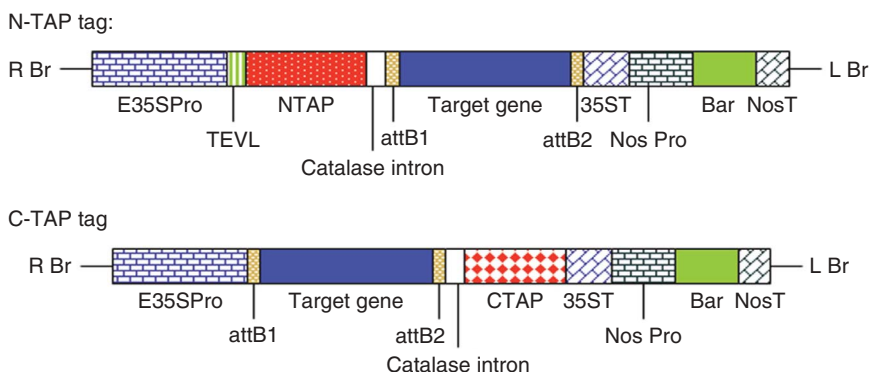


FIGURE 36.3 Structure of the T-DNA expressing TAP-tagged target protein. The ECaMV 35S promoter (E35SPro) and CaMV 35S polyadenylation region (35SST) are used to express the TAP-tagged gene. TEVL in N-TAP tag figure designates the TEV untranslated leader. R Br and L Br represent right and left borders, respectively. See page 530 for text discussion of this figure.

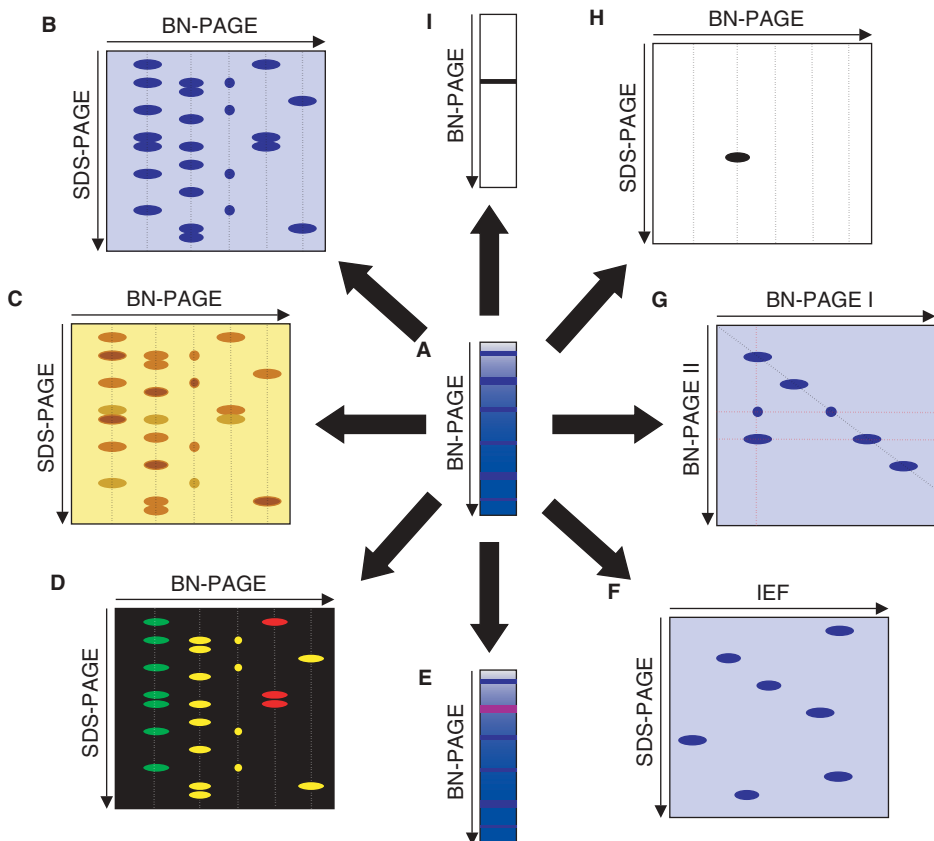


FIGURE 38.2 Schematic illustration of common downstream applications of 1D BN-PAGE. **(A)** 1D BN PAGE, unstained, **(B)** BN/SDS-PAGE for the resolution of protein complex subunits. CBB stained. **(C)** BN/SDS-PAGE for the resolution of protein complex subunits. Silver stained. **(D)** BN/SDS PAGE for the resolution of protein complex subunits. DIGE, pre-PAGE labeling of the proteins using different fluorophores for two different samples. Subunits depicted in yellow are equally abundant in both samples. Those displayed in red or green are more abundant in sample 1 or 2. **(E)** In-gel activity stain of a 1D BN-PAGE, also possible after 2D BN/BN-PAGE. **(F)** 3D BN/IEF/SDS-PAGE of a single band cut out from a 1D BN-PAGE. Proteins have been electro-eluted prior to IEF. CBB staining. **(G)** 2D BN/BN-PAGE. CBB stained. **(H)** Immunoblot of a 2D BN/SDS-PAGE. **(I)** Immunoblot of a 1D BN-PAGE. See page 564 for text discussion of this figure.